COVID-19 Twitter Monitor: Aggregating and visualizing COVID-19 related trends in social media

Joseph Cornelius[*] **Tilia Ellendorff**[†‡] **Lenz Furrer**[†‡] **Fabio Rinaldi**[*†‡]

[*]Dalle Molle Institute for Artificial Intelligence Research (IDSIA)
[†]Swiss Institute of Bioinformatics
[‡]University of Zurich, Department of Computational Linguistics
{joseph.cornelius,fabio.rinaldi}@idsia.ch
{tilia.ellendorff, lenz.furrer}@uzh.ch

Abstract

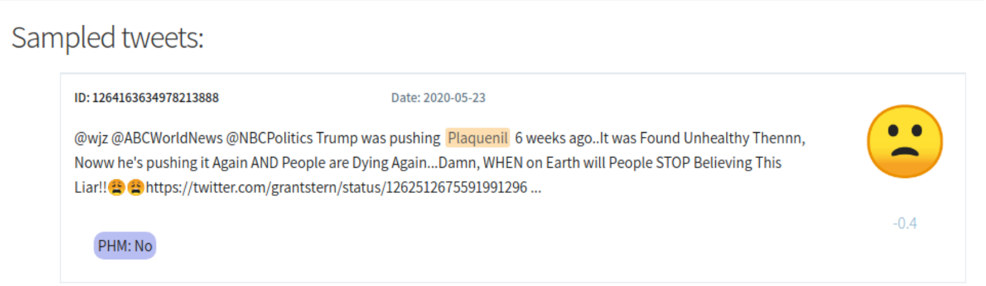
Social media platforms offer extensive information about the development of the COVID-19 pandemic and the current state of public health. In recent years, the Natural Language Processing community has developed a variety of methods to extract health-related information from posts on social media platforms. In order for these techniques to be used by a broad public, they must be aggregated and presented in a user-friendly way. We have aggregated ten methods to analyze tweets related to the COVID-19 pandemic, and present interactive visualizations of the results on our online platform, the COVID-19 Twitter Monitor. In the current version of our platform, we offer distinct methods for the inspection of the dataset, at different levels: corpus-wide, single post, and spans within each post. Besides, we allow the combination of different methods to enable a more selective acquisition of knowledge. Through the visual and interactive combination of various methods, interconnections in the different outputs can be revealed.

1 Introduction

Today, social media platforms are important sources of information, with a usage rate of over 70% for adults in the USA and a continuous increase in popularity.[1] Platforms like Twitter are characterized by their thematic diversity and realtime coverage of worldwide events. As a result, social media platforms offer information, especially about ongoing events that are locally and dynamically fluctuating, such as the SARS-CoV-2 (COVID-19) pandemic. The micro-posts contain not only the latest news or announcements of new medical findings but also reports about personal well-being and the sentiment towards public health interventions. Information contained in tweets has multiple usages. For instance, domain experts could discover how recent scientific studies find an echo in social media, while the health industry could find patients' reactions to certain drugs, and the general public could observe trends towards popular topics (e.g. the alternating popularity of politicians over time).

To provide universal access to the acquisition of knowledge from public discourse about COVID-19 on Twitter, social media mining methods have to be easily applicable.

[1] https://www.pewresearch.org/internet/fact-sheet/social-media/



Sampled tweets:

ID: 1264163634978213888 Date: 2020-05-23

@wjz @ABCWorldNews @NBCPolitics Trump was pushing Plaquenil 6 weeks ago..It was Found Unhealthy Thennn, Noww he's pushing it Again AND People are Dying Again...Damn, WHEN on Earth will People STOP Believing This Liar!!😩😩https://twitter.com/grantstern/status/1262512675591991296 ...

-0.4

PHM: No

Figure 1: A single COVID-19 related tweet annotated with our drug brand name detection system and classified by our systems for personal health mention (PHM) identification and sentiment analysis.

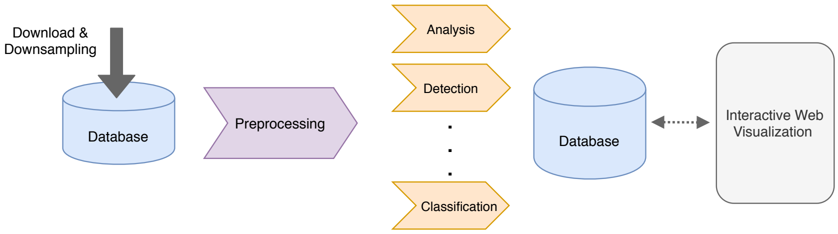


Figure 2: A schematic of the data acquisition and processing pipeline of our web platform.

We present a platform called COVID-19 Twitter Monitor that analyzes, categorizes, and extracts textual information from posts on Twitter. Our platform provides insights into the development and spread of the COVID-19 pandemic. The information is retrieved at different levels: spans within each post (e.g. drug name detection), single post (e.g. text language identification) and corpus level (e.g. distribution of hashtags). We use a variety of different methods to analyse pandemic-related micro-posts with a focus on health issues.

In contrast to existing platforms, we aim to provide a platform for newly emerging diseases for which no disease-specific supervised datasets are available. Therefore, our approach concentrates on unsupervised and pre-trained supervised methods that can be expected to generalize to unseen conditions. We have created an interactive web interface that integrates these methods and thus simplifies the use of social media mining methods. When visualizing the results, different methods are combined to display underlying connections in the extracted information. With this approach, we hope to make the contained information more accessible and provide an increased level of insight. Since we present the first version of our web platform[2], this is only an initial description of our approach and the preliminary results.

2 Related Work

Numerous studies show that social media platforms are particularly suitable for obtaining health-related information like the identification of drug abuse or the tracking of infectious diseases (Woo et al., 2016; Paul et al., 2016; McGough et al., 2017; Cocos et al., 2017). The detection of infectious disease outbreaks by using data from Twitter has already been tested, for instance in the case of the Ebola virus (Odlum and Yoon, 2015) and the Zika virus (Juric et al., 2017; Mamidi et al., 2019). Paul et al. (2016) propose using social media data for disease surveillance by locally determining and forecasting influenza prevalence. They address obstacles that the typically informal text on social media platforms presents to common language processing systems, and stress the need for normalized medical terminology and general linguistic normalization for social media. In addition, they argue that sentiment analysis is particularly suitable for analyzing the patient's impression of the quality of health care or public opinion on certain drugs. Likewise, Mamidi et al. (2019) apply sentiment analysis to tweets concerning virus outbreaks to gain information for disease control.

Joshi et al. (2019) describe how to identify disease outbreaks using text-based data. Like Wang et al. (2014), they show how topic modeling can be used to discover health-related topics in the micro-posts. Twitter adapted topic models were also utilized by Paul and Dredze (2011) to conduct syndromic surveillance regarding the flu.

Lamb et al. (2013) showed that more intricate data retrieval techniques can be applied to noisy Twitter data, such as identifying posts that mention personal health (PHM). Further fine-grained analyses were conducted by them, for example, to determine whether the tweet concerns the health of the author or a person familiar with the author. Going even further, Yin et al. (2015) showed that Twitter-related PHM classifiers can operate disease agnostic; they trained their system on four disease contexts and could achieve a precision of 0.77 on 34 other health issues.

Several platforms have already explored using social media data for health surveillance and disease tracking. Lee et al. (2013) uses Twitter data for influenza and cancer surveillance but focuses on the computation of various distributions, e.g. the temporal change in quantity or the most frequent words

[2] `https://covid19smm.nlp.idsia.ch`

	total number	unique	tweets containing
URLs	372228	283887	323760 *(72.76%)*
Domains	372228	37282	323760 *(72.76%)*
Hashtags	380744	93454	136279 *(30.63%)*
Languages	425721	53	425721 *(95.67%)*
RxNorm Brand Names	1285	101	1218 *(0.27%)*
Preprint Paper	152	124	149 *(0.03%)*
Tweets	444990	444990	–

Table 1: Statistics of the data collection with formerly 500K tweets after similarity filtering. For each feature, the total number of occurrences, the number of distinct occurrences, and the total number of tweets containing the feature are displayed.

of influenza tweets. In contrast, HealthMap[3] (Brownstein et al., 2008) lets you explore various disease outbreaks at the corpus and post level, but primarily uses google news as a data source. Systems such as Flutrack[4] (Chorianopoulos and Talvis, 2016) and Influenza Observations and Forecast[5] focus instead on map-based identification of infection disease outbreaks.

3 Data

We use the COVID-19 Twitter dataset published by the Panacea Lab (Banda et al., 2020). The dataset consists of daily Twitter chatter (about 4M tweets per day), collected by the Twitter Stream API for the keywords “COVD19”, “CoronavirusPandemic”, “COVID-19”, “2019nCOV”, “CoronaOutbreak”, “coronavirus”, and “WuhanVirus”. In addition, the Panacea Lab extended the dataset over time with Twitter datasets of research groups from the Universitat Autònoma de Barcelona, the National Research University Higher School of Economics, and the Kazan Federal University. As of June 14, 2020, version 14.0 of the dataset contains over 400 million unique tweets. Since the data has been collected irrespective of language, all languages are covered, with a pronounced prevalence of English (60%), Spanish (16%), and French (4%) tweets. An update is provided every two days and a cumulative update is provided every week. We use the cleaned version of the dataset provided by the Panacea Lab with all retweets filtered out.

4 Methods and Results

In the following section we describe the different methods used by our platform. We start by creating a randomly selected subset of the Panacea Lab’s COVID-19 Twitter dataset.[6] We then process this subset with the preprocessing steps described below and apply the methods for the data analysis as described in this section. Subsequently, we store the obtained results and the original data subset in a database accessed by our interactive web platform for visualization of the results, as illustrated in Figure 2.

4.1 Pre-Processing

For tweets used for topic modeling, sentiment analysis, and PHM identification, our system applies the following preprocessing methods:

- Without splitting the sentences, all tweets are tokenized using the spaCy[7] tokenizer.
- URLs are reduced to their domain names.
- The hash symbol “#” is removed from all hashtags.
- Multiple whitespace characters are removed.

[3] https://healthmap.org

[4] https://www.flutrack.org

[5] http://cpid.iri.columbia.edu

[6] This step is made necessary by the limitations of the computational infrastructure at our disposal for this activity.

[7] https://spacy.io/api/tokenizer

- Numbers are substituted with the token NUMBER.
- All @username are standardized to @USER.
- Camel-cased tokens are split into their components, e.g. “VirusOutbreak” to “Virus Outbreak”.
- Colloquial abbreviations such as “w/” for “with” are resolved.

4.2 Collection statistics

To reveal the following aspects in the dataset, we utilize distributions of multiple features. To identify different trending (sub)topics, we apply hashtag distributions and the domain/URL distribution to discover frequently-used sources and references of the tweet content. Furthermore, we use the distribution of languages in the dataset as a proxy to estimate the prevalence of the pandemic in different cultures and countries. To show the history of the intensity of discussion on different subtopics, we utilize the time distribution of tweets.

4.2.1 Hashtag Distribution

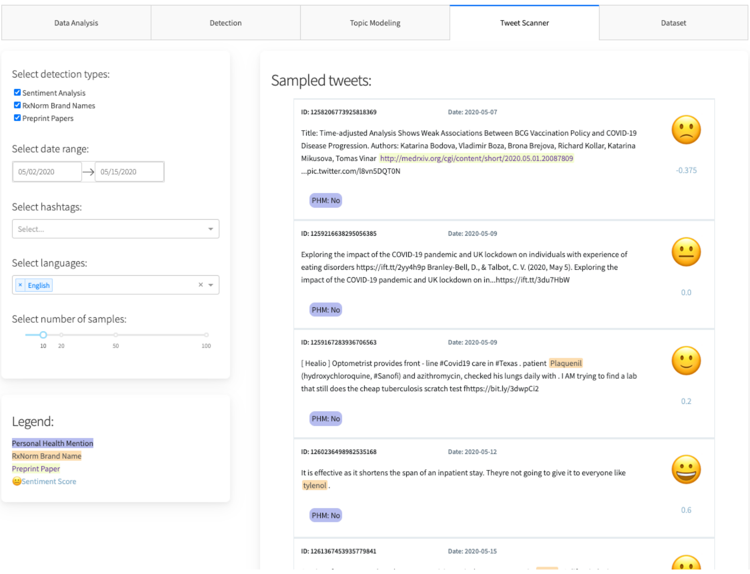
We extract all hashtags from each tweet in the corpus for the calculation of the hashtag distribution. In order to normalize the hashtags, we separate the hash symbol (“#”) from the hashtags, split all the hashtags written in camel-case into the individual parts and lowercase them. We utilize the resulting hashtags to compute the hashtag distribution. As shown in Table 1, with over 93K distinct hashtags in our subset with around 445K tweets, the data collection contains a broad diversity of subtopics.

4.2.2 Domain and URL Distribution

To detect URLs in tweets, we filter all posts for tokens that start with “http://”, “https://”, or “www.”. In addition, we use an unshortening method for each link, i.e. we retrieve the redirect information to get the target URL of links which have been compressed by the link shortening service (e.g. bitly.com). We derive the URL distribution from the processed URLs. By additionally reducing the URLs to their top-level domain (TLD) we get the domain distribution.

4.2.3 Language Distribution

For language detection, we use Nakatani Shuyo’s port to Python of the Language Detection Library for Java (Nakatani, 2010). The language detection is based on naive Bayesian filtering and is trained on



Select detection types:
☑ Sentiment Analysis
☑ RxNorm Brand Names
☑ Preprint Papers

Select date range:
05/02/2020 → 05/15/2020

Select hashtags:

Select languages:
English

Select number of samples:
10 20 50 100

Sampled tweets:

ID: 1258206773925818369 Date: 2020-05-07

Title: Time-adjusted Analysis Shows Weak Associations Between BCG Vaccination Policy and COVID-19 Disease Progression. Authors: Katarina Bodova, Vladimir Boza, Brona Brejova, Richard Kollar, Katarina Mikusova, Tomas Vinar http://medrxiv.org/cgi/content/short/2020.05.01.20087809 ...pic.twitter.com/l8vn5DQT0N

-0.375

PHM: No

ID: 1259216638295056385 Date: 2020-05-09

Exploring the impact of the COVID-19 pandemic and UK lockdown on individuals with experience of eating disorders https://ift.tt/2yy4h9p Branley-Bell, D., & Talbot, C. V. (2020, May 5). Exploring the impact of the COVID-19 pandemic and UK lockdown on in...https://ift.tt/3du7HbW

0.0

PHM: No

ID: 1259167283936706563 Date: 2020-05-09

[Healio] Optometrist provides front - line #Covid19 care in #Texas . patient Plaquenil (hydroxychloroquine, #Sanofi) and azithromycin, checked his lungs daily with . I AM trying to find a lab that still does the cheap tuberculosis scratch test fhttps://bit.ly/3dwpCi2

0.2

PHM: No

ID: 1260236498982535168 Date: 2020-05-12

It is effective as it shortens the span of an inpatient stay. Theyre not going to give it to everyone like tylenol .

0.6

PHM: No

ID: 1261367453935779841 Date: 2020-05-15

Figure 3: The screenshot of our web platform shows the subpage for the analysis of individual tweets. To obtain a selection of tweets, we filter them according to the analysis method, the date of creation, the contained hashtags, and their language. The selected tweets and the graphical indication of the different analysis results are displayed.

language profiles produced by Wikipedia abstracts. The classifier achieves an accuracy of 99% for 53 different languages. We produce the language distribution from the resulting language classification of each tweet. Our data collection contains tweets from all 53 languages, as can be seen in Table 1.

4.2.4 Temporal Tweet Distribution

We grouped the tweets by day and hour to display the temporal distribution and the temporal progression of tweets. Besides, we also allow the selection for tweets of a specific hashtag, see in Figure 5.

4.3 Topic Modeling

One of the hallmarks of Twitter is that an extensive discourse often accompanies emerging events. One method to find different themes within a vast amount of unlabeled posts is topic modeling. The goal of topic modeling is to discover in a probabilistic fashion distinct thematic structures in the underlying text corpus. It allows us to identify the main underlying topics that the collection of tweets is concerned with. We use the probabilistic, generative, and unsupervised Latent Dirichlet allocation (LDA) as a topic model (Blei et al., 2003; Hoffman et al., 2010).

We assume that each tweet contains a mixture of different topics in different proportions. The topics are considered to be a combination of different words. Since the number of topics is not known, we set the number of topics to twelve, a value we obtained by manual approximation. We utilize the LDA method of the python scikit-learn library.[8]

By means of two example topics found by the LDA method [“china”, “coronavirus”, “wuhan”, “spread”, “chinese”, “outbreak”,...] and [“mask”, “coronavirus”, “face”, “vaccine”, “reopen”, “wear”,...] it is possible to see how we can identify different thematic discussions within the tweets’ collection.

4.4 Drug Brand Name Detection

For the detection of drug brand names, we use the OntoGene Entity Recognizer (OGER) (Furrer and Rinaldi, 2017; Basaldella et al., 2017; Furrer et al., 2019). OGER is tailored for biomedical entity recognition and uses a combination of dictionary-based lookup and flexible matching. As dictionary, we use the RxNorm terminology,[9] a US-specific terminology that provides normalized names for all clinical drugs available on the US market. In the dictionary-based lookup for drug names, we consider only entries with the term types corresponding to brand names, since this allows us to significantly reduce the number of mismatches with a minor reduction of matches. In addition, we use a manually constructed blacklist with drug brand names that are ambiguous, such as “Android”, which is a mobile operating system as well as a brand name under which “methyltestosterone” is sold. To determine the precision of brand name detection, we manually evaluated 400 tweets containing tokens classified by our system as a brand name. The system achieves a precision of 0.67; however, 301 tweets are removed by prior filtering with the blacklist.[10]

4.5 Sentiment Analysis

Sentiment analysis is the interpretation and classification of emotions, used to differentiate the tweets into positive, negative, and neutral. We score the sentiment of each tweet in the range of [-1,+1] with -1 negative, 0 neutral, and +1 positive, by employing the Transformer based sentiment analysis model by Flair (Akbik et al., 2018). This model is trained on the IMDB review dataset (Maas et al., 2011) and achieves an accuracy of 98.87%. In addition to the sentiment analysis of individual tweets, we have an interactive panel to visualize the temporal variation of the sentiment score for tweets containing a selected hashtag. Furthermore, we calculate the average sentiment score over the entire period. This allows the user to track the sentiment of all tweets with specific hashtags, enabling the monitoring of changes in sentiment related to various topics in the public health discourse. For example, #mentalhealth

[8] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[9] https://www.nlm.nih.gov/research/umls/rxnorm/index.html

[10] The following words are included in the blacklist: “android”, “schiff”, “fml”, “pronto”, “alli”, “icar”, “cambia”, “ssd”, “tabloid”, “vanquish”, “ting”, and “propel”.

emerged on March 16th with a strongly negative sentiment four days after the first tweet with #lockdown occurred with slightly more positive sentiment in the data collection, one day after the WHO declared COVID-19 a pandemic.

4.6 Preprint Paper Detection

For the detection of URLs referring to preprint papers, we use the unshortened URLs of the tweets and extract the top-level domains (TLDs). We then compare them against a manually curated list of 55 TLDs of preprint servers. We can estimate the popularity of a preprint paper by the frequency of its mention in different tweets. Therefore we measure the string similarity[11] between all URLs to count and cluster URLs link to the same preprint paper. Our system reveals over 150 URLs of preprint paper in the subset of approx. 445K tweets, see Table 1.

4.7 Personal Health Mention Identification

In contrast to tweets that address general awareness of health issues, tweets about PHMs are concerned with the author's own health issues or those of persons familiar to the author. We use a BERT-based model for the identification of PHMs (Ellendorff et al., 2019). The model is trained to generalize to unseen health contexts. It was trained on two Twitter datasets belonging to two different flu-related contexts, flu infection and flu vaccination. The model was developed within the SMM4H shared task in 2019 and scored the best results. It achieved an overall accuracy of 0.877 and an F-score of 0.873 on the official test set, which includes tweets from seen and unseen health contexts.

4.8 Visualization

For the visualization of our interactive web interface, we use Dash[12]. We have currently limited the size of the dataset to 500K tweets to allow a prompt processing of the data and a seamless web visualization of the results. For a clearer presentation we display the calculation in five categorized tabs:

- **Data Statistics**: a tab that shows the distributions over the whole dataset, see Figure 6.
- **Detection**: a subpage for all detection methods that operate on the whole dataset (e.g. sentiment analysis, preprint paper identification, and medication brand name detection) as shown in Figure 4.
- **Topic Modeling**: a visualization of the results of the topic modeling using the pyLDAvis library.[13]
- **Tweet Scanner**: a subpage to show the detection methods on the tweet level, see Figure 3 and 1.
- **Dataset**: a panel with more detailed information about the tweets contained in the dataset (e.g. Tweet-ID and date).

5 Conclusion and Future Work

In this paper, we have presented the first version of our interactive web platform for the aggregation and visualization of different methods for social media mining regarding COVID-19.

Concerning social media mining, one of the biggest challenges in the analysis of emerging pandemics such as COVID-19 is that we initially have a very limited number of supervised datasets. Therefore, the selected methods applied to our platform are either unsupervised or have been previously trained on other social media datasets. We have demonstrated that by the combined visualization of different methods, underlying connections in the data can be revealed. In particular, we aimed to make the interactive web platform comprehensible for both a specialist audience and the general public.

As future work we intend to move the system to a more capable platform in order to be able to include a much larger number of tweets, and to add novel capabilities, such as the detection of tweets generated by bots as opposed to those generated by humans, through the usage of a tool such as the Botometer.[14]

[11] The string similarity is measured with the SequenceMatcher module `https://docs.python.org/3.6/library/difflib.html`

[12] `https://plotly.com/dash`

[13] `https://github.com/bmabey/pyLDAvis`

[14] `https://botometer.iuni.iu.edu/`

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration.

Marco Basaldella, Lenz Furrer, Carlo Tasso, and Fabio Rinaldi. 2017. Entity recognition in the biomedical domain using a hybrid approach. *Journal of Biomedical Semantics*, 8(1):51.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. 2008. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med*, 5(7):e151.

Konstantinos Chorianopoulos and Karolos Talvis. 2016. Flutrack.org: Open-source and linked data for epidemiology. *Health informatics journal*, 22(4):962–974.

Anne Cocos, Alexander G Fiks, and Aaron J Masino. 2017. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Tilia Ellendorff, Lenz Furrer, Nicola Colic, Noëmi Aepli, and Fabio Rinaldi. 2019. Approaching SMM4H with merged models and multi-task learning. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 58–61. Association for Computational Linguistics.

Lenz Furrer and Fabio Rinaldi. 2017. OGER: OntoGene's entity recogniser in the BeCalm TIPS task. In *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, pages 175–182.

Lenz Furrer, Anna Jancso, Nicola Colic, and Fabio Rinaldi. 2019. OGER++: hybrid multi-type entity recognition. *Journal of Cheminformatics*, 11(1):7.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre. 2019. Survey of Text-based Epidemic Intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6):1–19.

Radmila Juric, Inhwa Kim, Hemalatha Panneerselvam, and Igor Tesanovic. 2017. Analysis of Zika virus tweets: Could hadoop platform help in global health management? In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Alex Lamb, Michael Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.

Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using Twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1474–1477.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

Ravali Mamidi, Michele Miller, Tanvi Banerjee, William Romine, and Amit Sheth. 2019. Identifying key topics bearing negative sentiment on Twitter: insights concerning the 2015-2016 Zika epidemic. *JMIR Public Health and Surveillance*, 5(2):e11036.

Sarah F McGough, John S Brownstein, Jared B Hawkins, and Mauricio Santillana. 2017. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS neglected tropical diseases*, 11(1):e0005295.

Shuyo Nakatani. 2010. Language detection library for Java.

Michelle Odlum and Sunmoo Yoon. 2015. What can we learn about the Ebola outbreak from tweets? *American journal of infection control*, 43(6):563–571.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

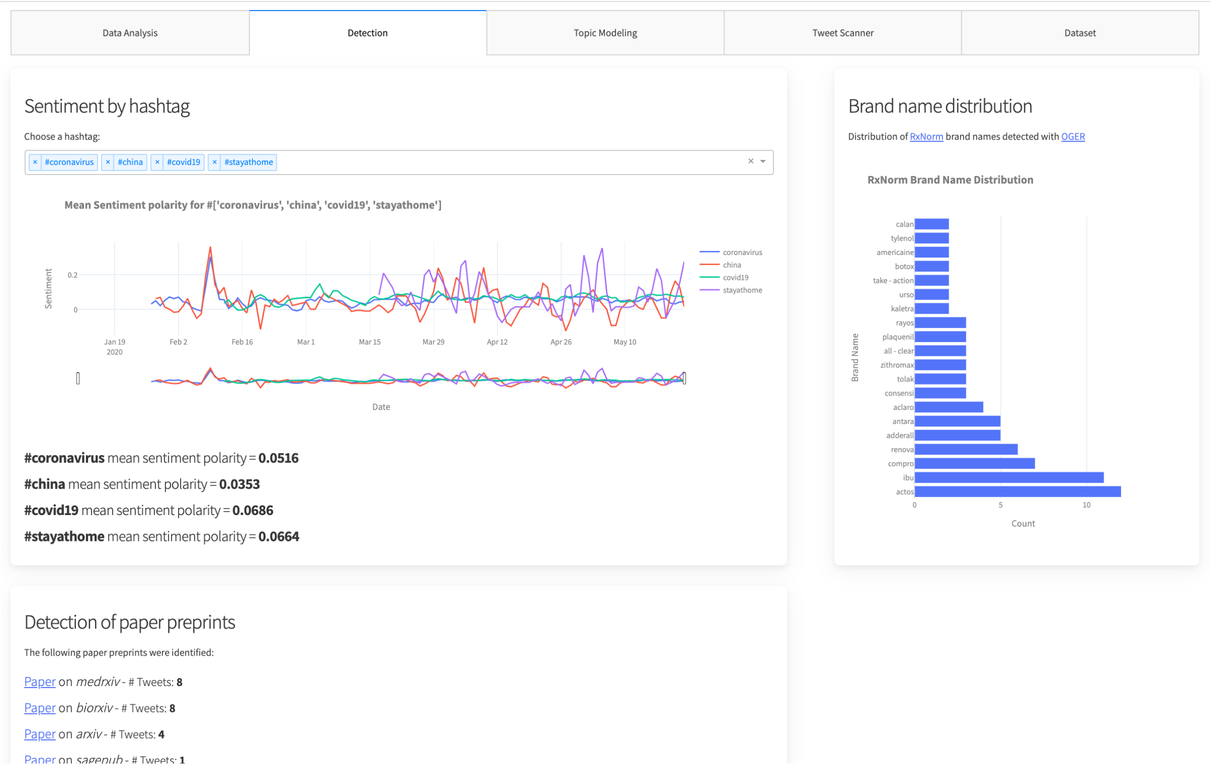
Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479.

Shiliang Wang, Michael J Paul, and Mark Dredze. 2014. Exploring health topics in chinese social media: An analysis of Sina Weibo. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Hyekyung Woo, Youngtae Cho, Eunyoung Shim, Jong-Koo Lee, Chang-Gun Lee, and Seong Hwan Kim. 2016. Estimating influenza outbreaks using both search engine query data and social media data in South Korea. *Journal of medical Internet research*, 18(7):e177.

Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. 2015. A scalable framework to detect personal health mentions on Twitter. *Journal of medical Internet research*, 17(6):e138.

A Appendices



Sentiment by hashtag

Choose a hashtag:

× #coronavirus × #china × #covid19 × #stayathome

#coronavirus mean sentiment polarity = **0.0516**

#china mean sentiment polarity = **0.0353**

#covid19 mean sentiment polarity = **0.0686**

#stayathome mean sentiment polarity = **0.0664**

Brand name distribution

Distribution of RxNorm brand names detected with OGER

Detection of paper preprints

The following paper preprints were identified:

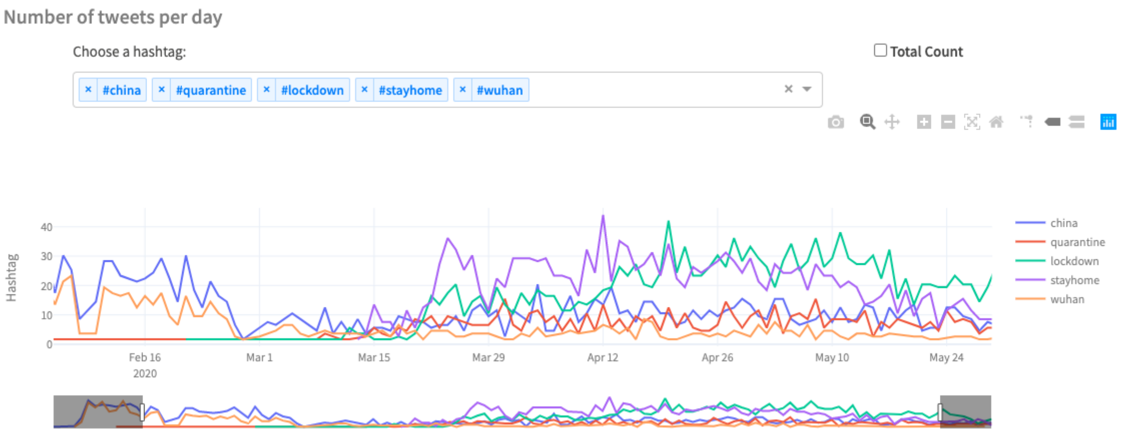
Paper on *medrxiv* - # Tweets: **8**

Paper on *biorxiv* - # Tweets: **8**

Paper on *arxiv* - # Tweets: **4**

Paper on *sagepub* - # Tweets: **1**

Figure 4: The screenshot of our web platform displays the subpage for the corpus-based detection methods. It shows the temporal sentiment analysis selected by hashtags, the detection of drug brand names, and the detection of preprint papers, respectively.



Number of tweets per day

Choose a hashtag: #china #quarantine #lockdown #stayhome #wuhan

□ Total Count

Figure 5: The detail from the screenshot of our web platform displays the temporal distribution of tweets. The progression can be visualized for all tweets as well as for tweets selected by their hashtags.

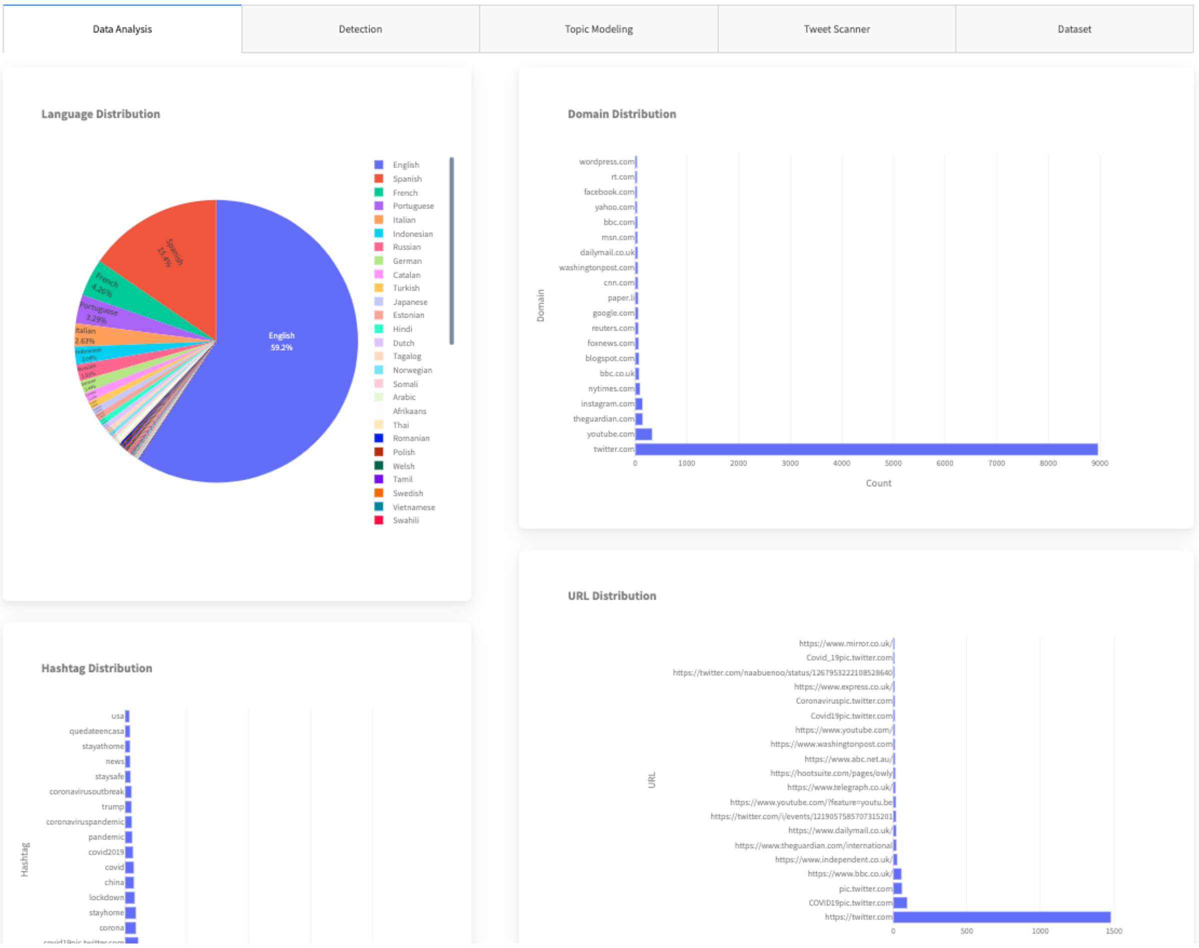


Figure 6: This screenshot of our web-platform shows the subpage for statistics of the dataset. The distribution of languages, domains, hashtags, and URLs are displayed.