Predicting COVID-19 cases using Reddit posts and other online resources

Felix Drinkall*, Janet B. Pierrehumbert*†
*Department of Engineering Science, University of Oxford
†Faculty of Linguistics, University of Oxford
felix.drinkall@lmh.ox.ac.uk

Abstract

This paper evaluates the ability to predict COVID-19 caseloads in local areas using the text of geographically specific subreddits, in conjunction with other features. The problem is constructed as a binary classification task on whether the caseload change exceeds a threshold or not. We find that including Reddit features, alongside other informative resources, improves the models' performance in predicting COVID-19 cases. On top of this, we show that exclusive use of Reddit features can act as a strong alternative data source for predicting a short-term rise in caseload due to its strong performance and the fact that it is readily available and updates instantaneously.

1 Introduction

A growing literature looks into the predictive power of social media (Evangelos et al., 2013). This predictive power has long been applied within quantitative finance (Xu and Cohen, 2018) and has been used to detect epidemics using the statistics of specific words associated with illness (Samaras et al., 2020). In this paper, we aim to use social media to predict the direction of the COVID-19 caseload in 4 local areas of the United States, using the state COVID-19 subreddits: Washington's r/CoronavirusWA, Florida's r/FloridaCoronavirus, California's r/CoronavirusCA and Texas' r/CoronaVirusTX. Data from the US was used due to its high level of activity on Reddit. Because the language was dynamic during the outbreak of this previously known virus, we did not follow (Samaras et al., 2020) in tracking keywords selected a priori, such as "Influenza". Instead, we used objective inclusion criteria to find which words were most predictive in each location.

To determine how well Reddit comments can predict future COVID-19 caseloads, this paper adopts the pipeline of Hofmann et al. (2020), a statistical NLP study using Reddit data in a very different application area (predicting the creation of new complex words). The pipeline uses a sliding window over the data stream, with each interval serving as the training data to predict the outcome in the subsequent window. Our model buckets every comment on the local subreddit into a set of daily documents F and selects a set W of important words using the inclusion criteria outlined in Section 2. The TF-IDF, T, of word w in the k^{th} document of F is calculated:

$$T_{w,F(k)} = tf_{w,F(k)} \times \log(\frac{|F|}{df_w}) \qquad (1)$$

Where $tf_{w,F(k)}$ is the number of occurrences of w in $F(k)$, $|F|$ is the number of documents in F and df_w is the number of documents that contain w. This statistic provides a good method for comparing how over-represented a word is in each document. Once the TF-IDF is calculated for all of W, we take the 7-day moving average (7-MA) of TF-IDF. The 7-MA is used throughout this study because of fluctuations in language usage according to the day of the week and because the caseload reports have artefacts from the day of the week. This time-series data is then tabulated and combined with other relevant datasets in Section 2 to determine which features are important for our prediction task.

2 Datasets and Predictors

The data that we seek to predict is provided by the COVID-19 Tracking Project (CTP) [1]. The current caseload (CCL) is also considered as a predictor for the subsequent change in caseload. Update frequency: 24 hours. Start date: 13/01/2020

The other predictors come from three sources. Each provides data about each day, but update speed differs. They were combined into a time-series dataset using data up to 17/01/2020. Days where data was incomplete were deleted.

Oxford COVID-19 Government Response Tracker (OxCGRT) - The OxCGRT (Hale et al., 2020) was used to identify which government measures were in place at each time. The data is structured into indicators covering a wide range of policies, including containment, health and economic measures, as well as an overall stringency score. Update frequency: "continuously", but due to human data collection, it can be variable; daily periodicity. Start date: 01/01/2020

Google's COVID-19 Community Mobility Reports (GCCMR) [2] - The GCCMR provided movement data within different areas such as parks, workplaces etc. The data has a high degree of geographic specificity. The movement statistic is relative to a benchmark taken between Jan. 3rd and Feb. 6th 2020. Update frequency: 2-3 days. Start date: 15/02/2020.

Pushshift API - The Pushshift API from Baumgartner et al. (2020) was used to the compile datasets of entire target subreddits. Update frequency: real-time. The post count P is considered as a predictor. For each subreddit, we also select **Feature words** by finding the most over-represented words, compared to a reference corpus R. To construct R, posts were randomly selected from S, the Unix time stamp of each was taken, and the following 100 posts from the whole of Reddit were downloaded. S and R were matched for the quantity of text at each time, as illustrated in Figure 1. The term frequency ratio between R and S was calculated, and the top 50 words were selected. To avoid over-reliance on rare words, the top 50 words with the highest term frequency in the top 1000 words in S were added for a total of 100 candidate word features. A chi-square test of independence was used to trim this candidate list to the 25 feature words with the most significant relationship to the target classes, and these were used in the prediction models. Appendix A lists the word features that were selected for each state. The important features are divided amongst named entities (locations, organisations, and people), technical terms, and terms referring to aspects of everyday life.

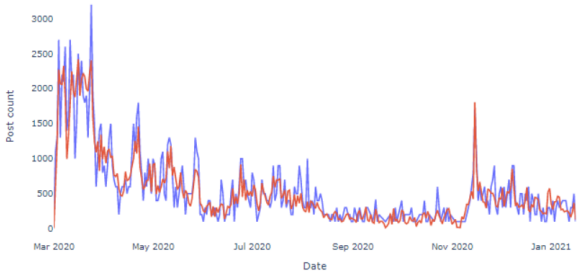

Figure 1: The post distribution of r/CoronavirusWA (S) and the matched post distribution (R).

3 Pipeline

The problem is presented as a classification task. For each day the features were tagged with a binary value that indicated whether the 7-MA of the caseload went up by more than a certain threshold value within the given time frame. We defined this threshold in two different ways, relative and absolute:

Relative change - the change is measured as a percentage of the current value.

$$\delta_r(t) = \frac{\mu(t+\tau) - \mu(t)}{\mu(t)}$$

Absolute change - the change is measured as the absolute change in the 7-MA.

$$\delta_a(t) = \mu(t+\tau) - \mu(t)$$

Where $\mu(t)$ is the 7-MA distribution of cases, and τ is the time delay that refers to the time horizon of the prediction. The relative threshold value ranges from 10% to 100%, and the absolute threshold ranges from 50 to 500. The predictive window iterates from 1 to 28 days.

Once the features were tagged with a binary value, the classes were balanced by identifying which class was larger and randomly deleting posts until they were equal in size. This was done to make the analysis more interpretable by making the

[1] https://covidtracking.com
[2] https://www.google.com/covid19/mobility/

accuracy comparable across the different classes. The minimum number of samples required was set at 40, 20 examples of both classes. Then the features were normalised, so they were all scaled between 0 and 1. Finally, the data was passed through the different classification models in Section 4.

4 Models

Tree-based models were used to determine the relative importance of the different features in predicting the changes in caseload. The performances of the models below were compared against a Support Vector Machine (SVM) with a linear kernel and a Logistic Classifier (LC) to see how the more complex models compared to classifiers with linear decision boundaries (Boser et al., 1992).

Random Forest (RF) - The RF model (Breiman, 2001) was chosen to show the viability of such a task. The benefit of using an RF model is that it decorrelates the different trees, which leads to robust results. The disadvantage is that it relies on a very dense feature set, which is problematic when the number of features grows to a size comparable to the number of samples.

Regularised Greedy Forest (RGF) - The RGF model (Johnson and Zhang, 2014) was chosen due to the added robustness from the fully-corrective regularized greedy search that learns the decision forests. This results in a sparse feature set by adopting L_1 and L_2 regularisation. We have compared this to an XGBoost model to deliver the best regularised model (Chen and Guestrin, 2016). This model also uses L_1 and L_2 regularisation to prevent overfitting.

5 Results

Each data source in Section 2 is used by itself to perform the classification task and then combined together to compare the performance across the different data sources. The default subreddit used below is Washington's r/CoronavirusWA since it has the largest subreddit by comment number.

Models (average)	7 days	14 days	21 days	28 days
RF (.828)	.805	.839	.838	.830
SVM (.814)	.766	.846	.829	.814
LC (.814)	.768	.830	.823	.833
RGF (.807)	.766	.834	.810	.819
XGBOOST (.785)	.753	.804	.798	.785

Table 1: This table shows the average performance across all relative thresholds at different prediction horizons using all of the data sources in Section 2.

Despite having more sparse feature sets, neither the RGF nor XGBOOST models outperform the RF model. The SVM and LC models perform very well; however, an RF model is used for the following analysis because of its high performance and because the Gini feature importances used in *sklearn.ensemble.RandomForestClassifier* are highly interpretable. The time delay that delivers the most precise results for each feature is tabulated in Table 2. In Table 2 the RF model is trained only using the data from each of the data sources in the Data Source column.

Data Source	7 days	14 days	21 days	28 days
all	.805	.839	.838	.830
T_w	.780	.808	.804	.791
M	.763	.808	.775	.801
G	.682	.751	.733	.748
P	.649	.663	.644	.648
CCL	.735	.622	.522	.417

Table 2: This table shows the average performance across all relative thresholds at different prediction horizons. The features are: T_w - the subreddit's word features; M - GCCMR movement data; G - OxCGRT government response data; P - daily post count; CCL - 7-MA of the current caseload.

Table 2 shows that for the majority of data sources, a 14 days prediction horizon yields the best results. Below, a delay of 14 days is used to compare the difference in performance using different thresholds. It also shows that T_w performs very well as a single feature class but that there is an improvement when all data sources are included.

Table 3 breaks down the performance across different thresholds for the increase. As found in many other studies, more extreme events are easier to predict. The highest performance, apart from using all the data sources, is found when only word features were used. These aspects of the performance are consistent across both the relative and absolute thresholds.

Feature importance - The importance of feature type in the RF model is tabulated in Table 4. The individual feature importances were added up by category. Clearly, the word features constitute the most important feature type in the prediction.

5.1 Generalising results

This section compares the performance in multiple states to see that these results are not unique for r/CoronavirusWA and that the framework is applicable within multiple regions. For this analysis, state subreddits that were comparable in size, and

	$\delta_r(t)$						$\delta_a(t)$					
m	0.2	0.4	0.6	0.8	1	$\mu + \sigma$	100	200	300	400	500	$\mu + \sigma$
all	.713	.853	.882	.875	.980	.839 + .079	.710	.794	.870	.957	.869	.849 + .087
T_w	.721	.813	.803	.889	.943	.808 + .075	.686	.750	.892	.886	.900	.824 + .085
M	.703	.814	.800	.843	.903	.808 + .076	.729	.710	.838	.700	.743	.737 + .055
G	.673	.693	.770	.771	.860	.751 + .057	.710	.542	.726	.714	.750	.689 + .069
P	.639	.614	.720	.643	.710	.663 + .051	.674	.636	.520	.700	.648	.633 + .062
CCL	.414	.479	.560	.771	.870	.622 + .163	.436	.702	.767	.900	.923	.681 + .204
# cases	264	140	100	70	54		264	152	92	70	64	

Table 3: **Performance** across the varying thresholds using an RF model, at a 14-day prediction horizon. Features are the same as Table 2; # cases - the number of data points.

	$\delta_r(t)$						$\delta_a(t)$					
m	0.2	0.4	0.6	0.8	1	$\mu + \sigma$	100	200	300	400	500	$\mu + \sigma$
T_w	.488	.507	.563	.630	.632	.564 + .067	.490	.612	.784	.745	.762	.655 + .115
M	.287	.238	.305	.251	.281	.272 + .027	.290	.210	.104	.131	.116	.193 + .076
G	.175	.202	.109	.086	.070	.128 + .057	.183	.114	.051	.026	.043	.090 + .057
P	.011	.013	.013	.006	.002	.009 + .005	.011	.012	.023	.004	.008	.010 + .006
CCL	.035	.020	.012	.015	.013	.019 + .009	.016	.027	.029	.085	.066	.042 + .024

Table 4: **Feature importances** across varying thresholds using an RF model, at a 14-day prediction horizon. As in Table 3, the features are the same as in Table 2.

were culturally different from one another were used. The list of subreddits is shown in Table 5.

State	Subreddit	Start date	Comments
Washington	r/CoronavirusWA	01/03/20	178998
California	r/CoronavirusCA	01/03/20	138121
Texas	r/CoronaVirusTX	01/03/20	129739
Florida	r/FloridaCoronavirus	07/03/20	83221

Table 5: This table shows data from the US state subreddits used as of 17/01/2021.

The words used for a given prediction are included in the Appendix in Table 7. The biggest class is city and hospital names, and other notable classes are organisations, people and technical epidemiological language. Notably, the feature words are topically rather than semantically informative.

In Table 6, the comparison feature sets M, G and CCL are combined and compared against the performance using all features, and the Reddit features alone. It shows that there is an improvement in performance across all time periods when using Reddit data alongside other data sources. The improvement that the subreddit data provides is generally more significant for larger subreddits. The NR features generally outperform the T_w features; however, this is not surprising since the M and G feature sets provide high-quality complementary information that should result in a higher performance than using each feature set in isolation; the M features show whether the government restrictions outlined by the G features are being listened to. The fact that there is a performance improvement when using all of the features shows that the Reddit data provides further information that is not captured in the NR features. Reddit data provides an insight into what people are talking and thinking about, which could cause people to not listen to the government restrictions.

Using only T_w yields very good results at a 7-day prediction horizon relative to other variable sources. COVID-19 has an incubation period of 5-6 days (Yu et al., 2020), it also takes a between 24-48 hours to get a PCR test result (Larremore et al., 2021). This high relative performance at a 7-day time delay suggests that the language on Reddit indicates whether the population is contracting

	7 days				14 days				21 days				
	All	NR	T_w	Diff.	All	NR	T_w	Diff.	All	NR	T_w	Diff.	Av. diff.
Washington	.805	.783	.780	.023	.839	.821	.808	.019	.838	.807	.804	.031	.024
California	.765	.722	.798	.043	.730	.740	.704	-.010	.808	.784	.735	.024	.019
Texas	.807	.781	.774	.027	.783	.783	.705	.000	.760	.772	.707	-.012	.005
Florida	.850	.851	.833	-.002	.859	.852	.804	.008	.850	.840	.751	.010	.006
Average	.807	.784	.803	.023	.803	.799	.755	.004	.814	.801	.749	.013	.014

Table 6: Performance across different states using an RF model. This is the average accuracy using different relative threshold values. NR - all data sources other than the subreddit data; T_w - as above; Diff. = All - NR

the disease more rapidly than other data sources do. That is, Reddit provides a strong live indicator of the experience and concerns in the population at any given time. In conjunction with real-time update frequency referenced in Section 2, this makes the use of subreddit data very convincing.

6 Conclusion

It is clear that the content of a local subreddit is a valuable data source for predicting the COVID-19 caseload in specific regions. The T_w features provided the best single feature set in almost all experimental setups, as seen in Table 3. When combined with the comparison feature sets in Table 6, the T_w features provided complementary information that resulted in a performance improvement. The results in Washington were also reproduced in other states, highlighting the robustness of the method used. A further advantage is that subreddit data is readily available. As is shown in Section 2, many of the other data sources take hours/days to update, and some only exist because the world is in a pandemic, as is the case with the GCCMR data.

There is also scope for future development using other machine learning techniques. In particular, using contextualised word embeddings has the potential to exploit the semantic relationships between words that are not well captured by a Bag-of-Words approach.

Acknowledgements

This work was supported in part by a grant from the Engineering and Physical Sciences Research Council (EP/T023333/1).

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Kalampokis Evangelos, Tambouris Efthimios, and Tarabanis Konstantinos. 2013. Understanding the predictive power of social media. *Internet Research*, 23, No. 5.

Thomas Hale, Sam Webster, Anna Petheric, Toby Phillips, and Beatriz Kira. 2020. Oxford covid-19 government response tracker. *Blavatnik School of Government.*

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. Predicting the growth of morphological families from social and linguistic factors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7273–7283, Online. Association for Computational Linguistics.

Rie Johnson and Tong Zhang. 2014. Learning nonlinear functions using regularized greedy forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):942–954.

Daniel B. Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M. Burke, James A. Hay, Milind Tambe, Michale J. Mina, and Roy Parker. 2021. Test sensitivity is secondary to frequency and turnaround time for covid-19 screening. *Science advances*.

Loukas Samaras, Elena García-Barriocanal, and Miguel-Angel Sicilia. 2020. Comparing social media and google to detect and predict severe epidemics. *Nature - Sci Rep 10.*

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Ping Yu, Jiang Zhu, Zhengdong Zhang, and Yingjun Han. 2020. A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period. *J Infect Dis.*

A Appendix

A.1 Feature words

State	Feature words	
Washington	Locations	'bothell', 'kirkland', 'omak', 'oroville', 'seatac', 'skagit', 'snohomish', 'spokane', 'thurston'
	People	'bedford', 'culp', 'inslee'
	Organisations	'esd', 'peuc'
	Technical	'7day', 'coronavirus', 'health', 'sick', 'virus'
	Other	'adjudication', 'business', 'news', 'open', 'places', 'social'
California	Locations	'alameda', 'huntington', 'merced', 'modesto', 'monterey', 'norcal', 'solano', 'sonoma', 'stanislaus', 'stockton'
	People	'garcetti'
	Organisations	'ihme'
	Technical	'cases', 'comorbidities', 'sick'
	Other	'aerosols', 'californian', 'californians', 'certain', 'city', 'defying', 'school', 'shelterinplace', 'state', 'states'
Texas	Locations	'abilene', 'brazoria', 'christi', 'galveston', 'houston', 'frisco', 'nueces'
	People	
	Organisations	'government'
	Technical	'antigen', 'c19', 'cases', 'death', 'hospitalizations', 'unmasked'
	Other	'city', 'counties', 'kids', 'week', 'weeks', 'woodlands'
Florida	Locations	'broward', 'duval', 'hillsborough', 'miamidade', 'pensacola', 'pinellas', 'sarasota'
	People	'deathsantis', 'desatan', 'gillum'
	Organisations	'arcgis', 'fdoh'
	Technical	'7day', 'cov19', 'kn95', 'mask', 'pandemic', 'vax'
	Other	'deathsentence', 'news', 'positivity', 'school', 'snowbirds', 'statewide', 'wear'

Table 7: Feature words for a 14-day prediction horizon at a relative threshold of 0.6.