Understanding Weekly COVID-19 Concerns through Dynamic Content-Specific LDA Topic Modeling

Mohammadzaman Zamani[1], **H. Andrew Schwartz**[1], **Johannes Eichstaedt**[2], **Sharath Chandra Guntuku**[3], **Adithya Virinchipuram Ganesan**[1], **Sean Clouston**[1], and **Salvatore Giorgi**[3]
[1] Stony Brook University [2] Stanford University [3] University of Pennsylvania
mzamani@cs.stonybrook.edu

Abstract

The novelty and global scale of the COVID-19 pandemic has lead to rapid societal changes in a short span of time. As government policy and health measures shift, public perceptions and concerns also change, an evolution documented within discourse on social media. We propose a dynamic content-specific LDA topic modeling technique that can help to identify different domains of COVID-specific discourse that can be used to track societal shifts in concerns or views. Our experiments show that these model-derived topics are more coherent than standard LDA topics, and also provide new features that are more helpful in prediction of COVID-19 related outcomes including mobility and unemployment rate.

1 Introduction

In early 2020, the entire world slowly became aware of a severe respiratory disease (Adhikari et al., 2020) with often catastrophic consequences (Zaim et al., 2020; Santesmasses et al., 2020). Since then, COVID-19 has infected millions of people worldwide and is on a trajectory to cause more than 1 million deaths globally before the end of 2020 (Medicine, 2020). Government and public health agencies have sought to mobilized and promote protective measures (Sen-Crowe et al., 2020) as scientific efforts rapidly build a foundation of knowledge to treat the disease, and economies have experienced rapid drops in employment (Zaim et al., 2020). These ongoing changes make the COVID-19 pandemic a time of immense and multifaceted social change that is rapidly evolving and largely, if imperfectly, narrated from millions of voices on social media.

User generated discourse like social media provide rich and continuous data on the evolution of significant world events - be it pandemics, hurricane relief, earthquakes, or wildfires (Thackeray et al., 2012). Given the rapidly evolving nature of such public health emergencies, the ability to extract and quantify the progression of topics can provide a unique window into the concurrent societal change. Realizing this objective is partially met through the goals of topic modeling, such as Latent Dirichlet Allocation (LDA; Blei et al. (2003)). However, tracking significant world events present 2 challenges to standard topic models: (1) the need for rapidly evolving topics rather than topics from a single snapshot of language at a time, and (2) the need to focus on event-relevant lexical patterns rather than general patterns.

Latent Dirichlet Allocation (LDA) is one of the most commonly used topic modeling methods, whereby probabilities of words belonging to topics (clusters of semantically related words) are derived from textual data. Not only can this provide a good set of features for predictive models (Brody and Elhadad, 2010; Zamani et al., 2018b), but also a tool for generating hypotheses and gaining insight in a manner easily interpretable by humans (Schwartz et al., 2013; Hu et al., 2012).

In this paper, we present and evaluate modifications to LDA, addressing the two aforementioned challenges, to focus on capturing topics that can characterize evolving interests specifically for COVID-19. Addressing such challenges enables several applications including monitoring the impact of a specific event on social, emotional, mental well-being and behaviours (Zamani et al., 2018a; Mirzaei et al., 2019). Building on previous work in online topic modeling (Canini et al., 2009), we propose the creation of short-interval *dynamic topics* that are updated on a regular basis (weekly or monthly). Furthermore, because words in social media posts cover a wide variety of domains (even when limited to posts containing COVID-19 keywords), we introduce a *content-specific* preprocessing step that focuses the lexicon on the domain. This limits the generation of incoherent or general domain topics — rather than a single general health

care topic, multiple health care topics may emerge, such as those related to vaccines and testing. This is especially important in the context of rapidly evolving public health emergencies like COVID-19, where conversations around physical distancing and protective measures have evolved as national policy has changed (Ross, 2020).

Our **contributions** include the formalization and evaluation of methods for topic modeling over (1) time series language data and (2) content focused lexical patterns; (3) open source data set of dynamic COVID-specific topics by week[1]; (4) demonstration that such topics can be used effectively as features to predict future US county-level mobility and unemployment.

2 Methods

2.1 Dynamic LDA topic modeling

LDA topic modeling estimates two sets of distributions: 1) representing each document as a multinomial distribution over T topics and 2) representing each topic with a multinomial distribution over W words. Blei, NG, and Jordan (2003) presented the LDA procedure to approximate the maximum-likelihood estimate for these distribution. Griffiths (Griffiths, 2002) presented a Gibbs Sampling based approach, which consists of a symmetric Dirichlet prior for both topics and words distributions followed by a Markov chain Monte Carlo inference. In this approach, at step i the topic assignment of word w_i is sampled according to the following conditional distribution:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(.)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,.}^{(d_i)} + T\alpha} \quad (1)$$

where, $\mathbf{w}$ is the data set consisting of words $w_1, ..., w_n$, in which each w_i belongs to a document d_i. Also, sub-index $-i$ indicates that the token at position i is disregarded in the calculation, $\mathbf{z}_{-i}$ is the topic probability distribution over words, $n_{-i,j}^{(w_i)}$ is the number of times word w_i is assigned to topic j, $n_{-i,j}^{(.)}$ is the total number of words assigned to topic j, $n_{-i,j}^{(d_i)}$ is the number of times a word in document d_i is assigned to topic j, and $n_{-i,.}^{(d_i)}$ is the total number of words in document d_i.

Inspired by these online Gibbs sampling approaches, we present a simple *dynamic topic modeling* for streams of data, when we are able to store a portion of data. The main difference with the aforementioned online sampler is that, in our scenario, we do not aim for estimating the topics distribution for the whole data set. In fact, we have a set of topics for each batch of data, and track changes over time as we receive a new batch of data. Since topics represent different concepts in the data set, we can trace how the existing concepts evolve, as well as discover vanishing or newly trending concepts.

In dynamic topic modeling for in-batch sampling we repeatedly use Gibbs sampling as in Equation 1 until we meet a stoppage criteria. However when it comes to cross-batch sampling we perform a more conservative sampling in order to transfer the topics posterior distribution obtained from the last batch into the topics prior distribution of the new batch. As shown in Equation 2, we use topics distribution over words from the previous batch but document distribution over topics from the current batch. This way we transfer the topics distribution over words to the current batch of data, while words in the same document still have a higher chance to be assigned to the same topics.

$$P(z_i^t = j|\mathbf{z}_{t-1}, \mathbf{w}_t) = \left(\frac{n_j^{(w_i)} + \beta}{n_j^{(.)} + W\beta}\right)_{t-1} \left(\frac{n_{i,j}^{(d_i)} + \alpha}{n_{i,.}^{(d_i)} + T\alpha}\right)_t \quad (2)$$

Here, subindex t and $t-1$ determine that $n_j^{(w_i)}$, $n_j^{(.)}$ and W are obtained from batch $t-1$ while $n_{i,j}^{(d_i)}$, $n_{i,.}^{(d_i)}$ and T are calculated from batch t.

2.2 Content-specific topic modeling

Content-specific (CS) topic modeling is a process for preparing text data for topic modeling, such that the desired topics are limited in thematic scope. In this paper we apply this method to automatically derive topics related to COVID-19, though this method has previously been applied to drinking (Giorgi et al., 2020) and diabetes tweets (Griffis et al., 2020). The entire process includes four steps: (1) tokenization and colocations; (2) identifying words most associated with our theme; (3) filtering documents to only the most discriminative words; and (4) the topic modeling process.

We start with a thematically coherent corpus, which in our case is a set of tweets containing COVID-19 keywords, from the streaming Twitter API: stayhome, stayathome, virus, coronovirus, coronavirus, covid, 2019-ncov, covd, outbreak,

[1] https://github.com/wwbp/weekly_covid_lda_topics

pandemic, corona, corono, washyourhand, handwashing, and sarscov (Guntuku et al., 2020).

We note that COVID-19 tweets contain language that may not be associated with COVID-19; for example, "RT" will appear in a large sample of our tweets, though this may or may not be associated with COVID-19. Thus, we would like to identify COVID-19 language and restrict our LDA process to that. As such, we take a random sample of COVID-19 tweets and find a matched set of tweets not related to our theme (i.e., tweets *without* COVID-19 keywords). We combine the COVID-19 and matched data, tokenize each document, and identify collocations (i.e., word phrases which are more common than the individual words within the phrase).

Next, we create a tweet level binary outcome: 1 if the tweet contains COVID-19 keywords and 0 otherwise, in other words, 1 for COVID-19 tweets and 0 for matched tweets. We then calculate a weighted log odds ratio, using an informative Dirichlet prior to estimate the difference in frequency of a word across two corpora (i.e., COVID-19 and matched tweets) (Monroe et al., 2008; Jurafsky et al., 2014). The prior shrinks word frequencies towards those of a large background corpus, while the z-score of the log odds ratio controls variance in word frequencies. Taking the tokens most associated with the binary outcome, we filter each document in our COVID-19 corpus to contain only these tokens, and run topic modeling over this filtered corpus.

3 Data

We built a COVID-19 corpus from publicly available Twitter data (March 12, 2020 to the end of June 2020; 2.6 million tweets), pulled from the streaming API using a set of COVID-19 keywords. We also took a sample of tweets pulled from the random 1% stream, which did not contain the COVID-19 keywords, as our matched tweets. We break this corpus into two sets of weekly and monthly data sets. The former, which is used for the coherence and real time tracking experiments, spans 5 weeks of tweets and contains 150,000 tweets per week for each of the COVID-19 and matched tweets. The latter, used for prediction experiment, spans 4 months and contains 500,000 tweets per month.

For the prediction task, we collect two monthly US county COVID-19 related outcomes: We use the SafeGraph Places Patterns[1] data, which contains aggregated and anonymized foot traffic data for 6 million points across the US. We use three months of mobility data from April to June, 2020. We also collect two months (April and May, 2020) of yearly-adjusted unemployment rates as reported by the US Bureau of Labor Statistics (BLS)[2].

4 Experiments

We perform three tasks to evaluate our proposed method. First, we consider weekly topic *coherence* as an intrinsic evaluation metric. Next, we manually compare the evolution of hand selected topics to real-world events for *real time tracking*. Finally, as an extrinsic evaluation metric, we use our topics as feature for predicting monthly mobility and unemployment.

For each of our three experiments, we create 40 weekly and monthly topics (for each LDA method) using the Java-based Mallet software package (McCallum, 2002), which implements Gibbs sampling (Gelfand and Smith, 1990). For each LDA method, parameters are kept constant ($\alpha = 5/N$, $\beta = 0.1$), where N is the number of desired topics. Additionally, stop words are removed and the number of unique tokens is constant at 8,000. All text pre-processing as well as the *prediction* task is done with the Python package DLATK (Schwartz et al., 2017).

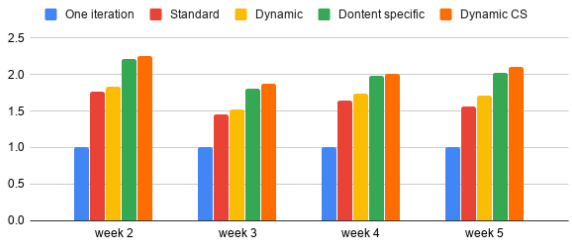

Figure 1: Normalized weekly point-wise mutual information coherence. Week 1 is excluded from this figure as dynamic CS and dynamic are the same as CS and standard, respectively, in Week 1.

Coherence We compare the average coherence (see appendix for formal definition) of topics across four LDA methods plus a baseline: standard LDA, CS, dynamic, and dynamic CS. For a topic t the

[1] https://www.safegraph.com/dashboard/covid19-commerce-patterns

[2] https://www.bls.gov/bls/newsrels.htm#OEUS

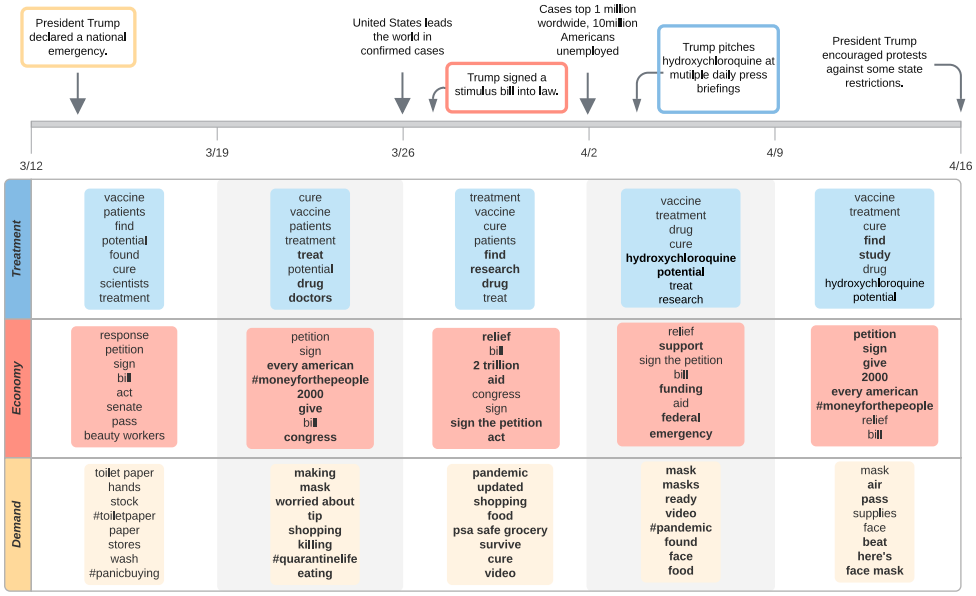

Figure 2: Timeline of COVID-19 events and topics over a five week period. Topics are visualized using the top 8 weighted words per week, bolded words are new to the top 8 topic words at that time point.

	m2→m3	m3→m4	m4→m5	Average
unigrams	0.15	0.20	0.34	0.23
unigrams + Standard LDA	0.19	0.25	0.43	0.29
unigrams + Dynamic	0.20	0.26	0.43	0.30
unigrams + CS	0.23	0.26	**0.46**	**0.32**
unigrams + Dynamic CS	**0.30**	0.26	**0.47**	**0.34**

(a) Mobility

	m2→m3	m3→m4	Average
unigrams	0.21	0.27	0.24
unigrams + Standard LDA	0.28	0.36	0.32
unigrams + Dynamic	0.27	**0.42**	0.34
unigrams + CS	0.24	0.38	0.31
unigrams + Dynamic CS	**0.31**	**0.40**	**0.36**

(b) Adjusted unemployment rate

Table 1: Prediction accuracies (Pearson r): language features from month n predict outcomes at month $n+1$. Bold indicates significant improvement over unigrams + standard method, according to paired t-tests.($p < .05$).

coherence score can be obtained as the summation of coherence score of each pair of words from the top selected words of that topic. For the coherence of pair of words there are various options, where we use point-wise mutual information (PMI) as suggested by (Newman et al., 2010):

$$Coherence(t) = \sum_{i<j} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (3)$$

where, w_i and w_j are from the top words of the topic t.

Our baseline is a single iteration of Gibbs sampling for standard LDA and is used for normalizing all scores.

Figure 1 shows the results of our *coherence* task. Not surprisingly, all four LDA methods outperform the single iteration baseline. We see a small increase in coherence when using the dynamic approach over standard LDA. Significant increases (above both the standard and dynamic approaches) occur when using the CS method, as determined by a t-test ($p < 0.05$). Finally, we see marginal increases in coherence (above the CS method) when using the the dynamic CS approach.

Real Time Tracking For the tracking task, the goal is to see if our topics are evolving alongside real world events. As such, we hand select three topics, labeled as *treatment*, *economy* and *demand* and show their evolution over a 5 weeks period. The timeline in Figure 2 shows the top eight weighted words within each topic, in addition to plotting notable COVID-19 events.

The *demand* topic starts with mentions of toilet paper and panic buying and moving to masks and general supplies. The *treatment* topic mentions hydroxychloroquine around the time the FDA approved emergency use (March 29) and when Trump was frequently mentioning the drug (April 3 through 5). Finally, the *economy* topic starts with calls for emergency federal help ("sign", "act", "pass") and then moves onto more specific information around the stimulus bill ("2 trillion" and "congress") before stabilizing.

Prediction Using a 10-fold cross validation setup, we predict two monthly COVID-19 related outcomes: mobility and unemployment rates. We extract monthly topic representations (standard, dynamic, CS and dynamic CS) for US counties who have written at least 7,500 words per month (i.e., each observation is a US county and is represented as a bag-of-topics).

Here, we use language features for a given month to predict the outcomes for the following month (e.g., topics from March are used to predict outcomes in April). For each month, we perform a 10-fold cross-validation of a regularized ridge regression and report the Pearson r between our predicted values and the actual rates. We use unigram features as the baseline and compared our gain by adding standard LDA features vs. dynamic CS LDA features. Results in Table 1 suggest that while topic features are adding on top of unigrams, the dynamic CS features perform significantly better.

5 Discussion

Discussions on social networks are rapidly changing, especially those centered around COVID-19. As such, the goal of the current study was to see if we could reasonably monitor those changes over time using a novel LDA topic-modeling approach, namely dynamic content-specific LDA. Results showed that we could reliably track topic evolution and leverage those dynamics to predict changes in real work outcomes (mobility and unemployment) We found that many such observations could be readily tracked, suggesting that we could reliably track discussions, even when the specific words or concerns used to reference that topic change.

This has a direct application to COVID-19 as it may help determine appropriate shifts in public health strategy as topics evolve towards consensus. Additionally, insofar as topics involve and reliably track symptoms, then evolution in reported symptoms may help in efforts to map out differences in strains or in disease evolution. It is also worthwhile to note that this model can be used to track *any* evolving discussion over time. Indeed, many discussions both in the field of public health or medicine and in other fields, such as understanding financial shifts or changes to public acceptance of known falsehoods or conspiracies, are discussed regularly online and may be tracked as they evolve. The evolution of such discussions can be telling, we think, and the mapping of that evolution warrants further interrogation.

A central goal in epidemiological modeling is monitoring psychiatric and health outcomes, and their correlates, as they shift over time. This study helps us to better identify and track the discourse around COVID-19 via an online social network and also helped to characterize changes in topics as they evolved over time. As such, LDA topic analysis identified and helped to characterize changing interests in both COVID-19 and its societal effects as they emerged over time. This study suggests that further research should seek to understand how these topics are associated with the distribution of COVID-19 cases and changes in mental health.

$$P(z_i = j|\mathbf{w}) * P(z_i = j|\mathbf{z})$$

(4)

A = P(z_i = $j|\mathbf{z}_t * P(z_i = j|\mathbf{w}_{t-1}) * P(z_i = j|\mathbf{z}_t)$

P($z_i = j|\mathbf{d}, \mathbf{z}) = P(z_i = j|\mathbf{d}) * P(z_i = j, \mathbf{d}|\mathbf{z})$

P($z_i = j|\mathbf{d}, \mathbf{z}) = P(z_i = j|\mathbf{d}) * P(z_i = j, \mathbf{d}|\mathbf{z})$(5)p(zi=j — d,z) = p(zi=j — d) * p(zi=j , d — z)

References

Sasmita Poudel Adhikari, Sha Meng, Yu-Ju Wu, Yu-Ping Mao, Rui-Xue Ye, Qing-Zhi Wang, Chang Sun, Sean Sylvia, Scott Rozelle, Hein Raat, et al. 2020. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (covid-19) during the early outbreak period: a scoping review. *Infectious diseases of poverty*, 9(1):1–12.

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

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 804–812.

Kevin Canini, Lei Shi, and Thomas Griffiths. 2009. Online inference of topics with latent dirichlet allocation. In *Artificial Intelligence and Statistics*, pages 65–72.

Alan E Gelfand and Adrian FM Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Salvatore Giorgi, David B Yaden, Johannes C Eichstaedt, Robert D Ashford, Anneke EK Buffone, H Andrew Schwartz, Lyle H Ungar, and Brenda Curtis. 2020. Cultural differences in tweeting about drinking across the us. *International Journal of Environmental Research and Public Health*, 17(4):1125.

Heather Griffis, David A Asch, H Andrew Schwartz, Lyle Ungar, Alison M Buttenheim, Frances K Barg, Nandita Mitra, and Raina M Merchant. 2020. Using social media to track geographic variability in language about diabetes: Infodemiology analysis. *JMIR diabetes*, 5(1):e14431.

Tom Griffiths. 2002. Gibbs sampling in the generative model of latent dirichlet allocation.

Sharath Chandra Guntuku, Garrick Sherman, Daniel C Stokes, Anish K Agarwal, Emily Seltzer, Raina M Merchant, and Lyle H Ungar. 2020. Tracking mental health and symptom mentions on twitter during covid-19. *Journal of general internal medicine*, pages 1–3.

Yuheng Hu, Ajita John, Fei Wang, and Subbarao Kambhampati. 2012. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI*, volume 12, pages 59–65.

Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit (2002).

Johns Hopkins University & Medicine. 2020. Coronavirus resource center.

Mehrdad Mirzaei, Shaghayegh Sahebi, and Peter Brusilovsky. 2019. Annotated examples and parameterized exercises: Analyzing students' behavior patterns. In *International Conference on Artificial Intelligence in Education*, pages 308–319. Springer.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224.

Katherine Ross. 2020. Why weren't we wearing masks from the beginning? dr. fauci explains. *TheStreet*.

Didac Santesmasses, José Pedro Castro, Aleksandr A Zenin, Anastasia V Shindyapina, Maxim V Gerashchenko, Bohan Zhang, Csaba Kerepesi, Sun Hee Yim, Peter O Fedichev, and Vadim N Gladyshev. 2020. Covid-19 is an emergent disease of aging. *MedRxiv*.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60.

Brendon Sen-Crowe, Mark McKenney, and Adel Elkbuli. 2020. Social distancing during the covid-19 pandemic: Staying home save lives. *The American journal of emergency medicine*.

Rosemary Thackeray, Brad L Neiger, Amanda K Smith, and Sarah B Van Wagenen. 2012. Adoption and use of social media among public health departments. *BMC public health*, 12(1):1–6.

Sevim Zaim, Jun Heng Chong, Vissagan Sankaranarayanan, and Amer Harky. 2020. Covid-19 and multi-organ response. *Current Problems in Cardiology*, page 100618.

Mohammadzaman Zamani, Anneke Buffone, and H Andrew Schwartz. 2018a. Predicting human trustfulness from facebook language. *arXiv preprint arXiv:1808.05668*.

Mohammadzaman Zamani, H Andrew Schwartz, Veronica E Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018b. Residualized factor adaptation for community social media prediction tasks. *arXiv preprint arXiv:1808.09479*.

6 Appendix

COVID-19 Keywords We used the following set of keywords to pull COVID-19 tweets from the streaming Twitter API: stayhome, stayathome, virus, coronovirus, coronavirus, covid, 2019-ncov, covd, outbreak, pandemic, corona, corono, washyourhand, handwashing, and sarscov.

LDA Parameters For each of our three experiments, we create 40 weekly and monthly topics (for each LDA method) using the Java-based Mallet software package (McCallum, 2002), which implements Gibbs sampling (Gelfand and Smith, 1990). For each LDA method, parameters are kept constant ($\alpha = 5/N$, $\beta = 0.1$), where N is the number of desired topics. Additionally, stop words are removed and the number of unique tokens is constant at 8,000. All text pre-processing as well as the *prediction* task is done with the Python package DLATK (Schwartz et al., 2017).

Coherence Definition For a topic t the coherence score can be obtained as the summation of coherence score of each pair of words from the top selected words of that topic. For the coherence of pair of words there are various options, where we use pointwise mutual information (PMI) as suggested by (Newman et al., 2010):

$$Coherence(t) = \sum_{i<j} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (6)$$

where, w_i and w_j are from the top words of the topic t.