Characterization of Potential Drug Treatments for COVID-19 using Twitter

Ramya Tekumalla¹ and Juan M Banda¹

¹ Georgia State University, Atlanta, Georgia, USA

BioHackathon series:

COVID-19 BioHackathon

Virtual conference 2020

Submitted: 25 May 2020

License

Authors retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](https://creativecommons.org/licenses/by/4.0/)).

Published by [BioHackrXiv.org](https://biohackxiv.org)

Abstract

Since the classification of COVID-19 as a global pandemic, there have been many attempts to treat and contain the virus. Although there is no specific antiviral treatment recommended for COVID-19, there are several drugs that can potentially help with symptoms. In this work, we mined a large twitter dataset of 280 million tweets of COVID-19 chatter to identify discourse around potential treatments. While seemingly a straightforward task, due to the informal nature of language use in Twitter, we demonstrate the need of machine learning methods to aid in this task. By applying these methods we are able to recover almost 21% additional data than with traditional methods.

Introduction

The World Health Organization (WHO) defines the Coronavirus disease (COVID-19) as an infectious disease caused by a newly discovered coronavirus and declared it a pandemic on March 11, 2020 (Organization, 2020). As of May 11, 2020, 4,148,034 cases were confirmed worldwide with 284,124 deaths and 1,428,161 cases recovered. Social media platforms like Twitter and Reddit contain an abundance of text data that can be utilized for research. Over the last decade, Twitter has proven to be a valuable resource during disasters for many-to-many crisis communication (Earle, 2010) (Zou, Lam, Cai, & Qiang, 2010) (Alam, Ofli, Imran, & Aupetit, 2018). Recently, Several works (Gao, Tian, & Yang, 2020) (Lu, 2020) (Sanders, Monogue, Jodlowski, & Cutrell, 2020) have provided insights on the treatment options and drug usages for COVID-19.

We utilized the largest available COVID-19 dataset (Banda et al., 2020) curated using a Social Media Mining Toolkit (Tekumalla & Banda, 2020b). This dataset consists of tweets related to COVID-19 from January 1, 2020. Version 9 was utilized for our experiments since it was the latest version at the time of experiments. We annotated ~283 million tweets using the drug dictionary (Tekumalla, Asl, & Banda, 2019) compiled from RxNorm (Medicine, n.d.) with 19,643 terms and validated in (Tekumalla et al., 2019) and (Tekumalla & Banda, 2020a).

Methods

In order to identify drug-specific tweets related to COVID-19, we applied the drug dictionary on the clean version of the dataset. With the presence of retweets, any signal found would be greatly amplified and hence only unique tweets must be utilized to identify the signals. The cleaned version of the dataset consists of 62 million unique tweets with no retweets. The Spacy annotator utility from the Social Media Mining Toolkit (Tekumalla & Banda, 2020b) was utilized to annotate the clean tweets. A total of 677,729 tweets were identified as tweets containing one or more drug terms in them. However, complex medical terms,

such as medication and disease names, are often misspelled by users on Twitter (Karimi, Metke-Jimenez, Kemp, & Wang, 2015). Using only the keywords with correct spellings leads to a loss of potentially important data, particularly for terms with difficult spellings. Some misspellings also occur more frequently than others, implying that they are more important for data collection and concept detection. To determine the misspelled drug tweets, we initially identified the top 200 drug terms from 677,729 drug tweets from the clean dataset to generate misspellings.

We employed two different methodologies, a machine learning approach called QMisSpell (Sarker & Gonzalez-Hernandez, 2018) and a more traditional approach of generating misspellings using keyboard layout distance. The RedMed (Lavertu & Altman, 2019) model's word embeddings was utilized in QMisSpell to generate misspellings. QMisSpell relies on a dense vector model learned from large, unlabeled text, which identifies semantically close terms to the original keyword, followed by the filtering of terms that are lexically dissimilar beyond a given threshold. The process is recursive and converges when no new terms similar (lexically and semantically) to the original keyword are found. A total of 2,056 unique misspelled terms were generated using QMisSpell. We examined all the misspellings and eliminated several terms (example: heroes, heroine generated from the original keyword heroin) which are common terms in English vocabulary and are not misspellings of drug terms. Post elimination, a total of 1,908 terms are identified as the misspelt terms.

The second methodology utilizes keyboard layout distance for generating the misspellings. For each term, each alphabet is replaced with the closest alphabet on the QWERTY keyboard. This process is recursive and ceases when it has looped through every alphabet in the term. For example, the term cocaine has 68 misspelled terms which vary from xocaine to cocaint. After eliminating the common vocabulary and duplicates, a total of 15,696 terms are identified as misspelled terms for this methodology. We annotated the clean dataset utilizing the misspelled terms from the methodologies. Table 1 summarizes the results. Overlap between QMisSpell and the Keyboard Layout generator is 4.9%. 33 terms were common between the two misspelled dictionaries.

Table 1. Term annotation results

Dictionary	Total number of terms	Total annotated terms
RxNorm drug Dictionary	19,643	768,752
QMisSpell	1,908	97,776
Keyboard Layout Method	15,696	76,394
Total	37,214	942,922

Table 2. Most frequent drugs found using the drug dictionary.

Drug Name	Frequency
hydroxychloroquine	91,696
chloroquine	66,904
agar	26,748
oxygen	20,214
zinc	13,235
vitamin d	11,770
azithromycin	9,992
nicotine	7,140
ibuprofen	6,804
dettol	6,779

Results

Not unexpectedly, Hydroxychloroquine was the most tweeted drug found on the dataset with chloroquine coming in as close second, as shown on Table 2. Interestingly enough, hydroxychloroquine was not the most misspelled word, chloroquine was, as we can see on Table 3.

Table 3. Most frequent misspelled drugs found using the two methods

Keyboard Layout Method Drug Name	Keyboard Layout Method Frequency	QMisSpell Method Drug Name	QMisSpell Method Frequency
chloroquine	16,361	hydroxychloroquine	6,650
vitamin a	12,094	allegra	6,266
agar	8,933	meted	5,135
hydroxychloroquine	2,878	ibuprofen	4,249
doral	2,445	propane	4,233
nicotine	2,445	coconut oil	2,228
cocaine	900	azithromycin	1,624
zinc	858	acetaminophen	726
septa	538	aluminum	672
aleve	487	chlorine	555

In our exploration of misspellings, we had 23 potential different spellings for Hydroxychloroquine. We have 78 possible misspellings from chloroquine as well. While not all possible misspellings are found in the actual Twitter data, a good combination of them (45%) are actually present, indicating how tricky it is to work with Twitter text data. The total added term counts can be found in Figure 1. As we can see in both Figure 1 and Table 1, by considering the possibility of misspellings, we find around 20% more data. When dealing with limited text availability, particularly drug terms in Twitter, it is vital to use these types of methods to recover as much signal as we can. Terms such as Agar, Coconut Oil and Zinc acquired high counts as listed on Table 3. This is due to a few irrelevant terms prevailing in the dictionary. The drug dictionary curated from RxNorm consists of terms from different semantic types (Ingredients, Semantic Clinical Drug Component, Semantic Branded Drug Component and Semantic Branded Drug). The Ingredients semantic type consists of different terms that are technically not drugs but are utilized in creating drugs (Eg: Gelatin, Agar, Coconut Oil, Zinc). Such terms were still included in the dictionary due to the granularity they bring to the research.

Frequency vs Drug Name

Figure 1: DrugTermOccurrences

Figure 1. Drug Term occurrences during COVID-19

Future Work

In this research, we generated misspellings utilizing two different methods. While the keyboard layout method is an automated method to generate misspellings, the QMisSpell method generates misspellings based on language models. Unfortunately, not many pre-trained models related to drug research are readily available. The language model used in this research was able to generate misspellings for 80% of the top 200 terms. It is immensely difficult to obtain drug related data to create a language model to generate misspellings for all terms due to unavailability of the data. In future, we would like to implement an auto correct module, which would rectify the misspellings at the text level. This would potentially demonstrate the trade off between generating misspellings vs auto correcting the spellings at the text level.

Conclusion

This research utilized the largest available COVID-19 dataset to identify the drugs that circumscribe around coronavirus. By employing two different spelling modules, we were able to recover 21% additional data which can be utilized for research. Using sentiment analysis, we can also identify how users respond to these drugs and identify the drugs that help with some symptoms. With plenty of careful research, this data can be utilized for extracting treatment options.

GitHub repository

- Link to the github repository - https://github.com/thepanacealab/covid19_biohackathon/tree/master/treatments_misinformation

Acknowledgements

This research was developed during COVID-19 Biohackathon April 5-11 2020. We thank the organizers for coordinating the virtual hackathon during the COVID-19 crisis.

References

- Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *arXiv*. Retrieved from <http://arxiv.org/abs/1805.05144>
- Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration. *arXiv*. doi:[10.5281/zenodo.3723940](https://doi.org/10.5281/zenodo.3723940)
- Earle, P. (2010). Earthquake Twitter. *Nature geoscience*, 3(4), 221–222. doi:[10.1038/ngeo832](https://doi.org/10.1038/ngeo832)
- Gao, J., Tian, Z., & Yang, X. (2020). Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *Bioscience trends*, 14(1), 72–73. doi:[10.5582/bst.2020.01047](https://doi.org/10.5582/bst.2020.01047)
- Karimi, S., Metke-Jimenez, A., Kemp, M., & Wang, C. (2015). Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55, 73–81. doi:[10.1016/j.jbi.2015.03.010](https://doi.org/10.1016/j.jbi.2015.03.010)
- Lavertu, A., & Altman, R. B. (2019). RedMed: Extending drug lexicons for social media applications. *Journal of biomedical informatics*, 99, 103307. doi:[10.1016/j.jbi.2019.103307](https://doi.org/10.1016/j.jbi.2019.103307)
- Lu, H. (2020). Drug treatment options for the 2019-new coronavirus (2019-nCoV). *Bioscience trends*, 14(1), 69–71. doi:[10.5582/bst.2020.01020](https://doi.org/10.5582/bst.2020.01020)
- Medicine, N. L. of. (n.d.). RxNorm [Internet]. Retrieved from <http://www.nlm.nih.gov/research/umls/rxnorm/>
- Organization, W. H. (2020). WHO characterizes COVID-19 as a pandemic. Retrieved from <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen>
- Sanders, J. M., Monogue, M. L., Jodlowski, T. Z., & Cutrell, J. B. (2020). Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19): A Review. *JAMA: the journal of the American Medical Association*. doi:[10.1001/jama.2020.6019](https://doi.org/10.1001/jama.2020.6019)
- Sarker, A., & Gonzalez-Hernandez, G. (2018). An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *Journal of biomedical informatics*, 88, 98–107. doi:[10.1016/j.jbi.2018.11.007](https://doi.org/10.1016/j.jbi.2018.11.007)
- Tekumalla, R., Asl, J. R., & Banda, J. M. (2019). Mining Archive.org's Twitter Stream Grab for Pharmacovigilance Research Gold. *bioRxiv*. Retrieved from <http://biorxiv.org/content/early/2019/12/03/859611.abstract>
- Tekumalla, R., & Banda, J. (2020a). A large-scale Twitter dataset for drug safety applications mined from publicly existing resources. *arXiv*. doi:[10.5281/zenodo.3606863](https://doi.org/10.5281/zenodo.3606863)
- Tekumalla, R., & Banda, J. M. (2020b). Social Media Mining Toolkit (SMMT). *arXiv*. Retrieved from <http://arxiv.org/abs/2003.13894>
- Zou, L., Lam, N. S. N., Cai, H., & Qiang, Y. (2010). Mining Twitter Data for Improved Understanding of Disaster Resilience. *Annals of the Association of American Geographers. Association of American Geographers*, 108(5), 1422–1441. doi:[10.1080/24694452.2017.1421897](https://doi.org/10.1080/24694452.2017.1421897)