Analyzing the impact of lockdown on the public sentiment: A Case Study

Abstract

Understanding the psychological impact of a pandemic is an important part of studying the dynamics of such an outbreak, and its impact on the world. This paper puts forward a case study that demonstrates the effect of the COVID-19 pandemic on the public sentiment, with a special focus on the Indian subcontinent. The study is based on a pipeline that involves preprocessing, visualization, natural language processing and statistical analysis of Twitter data in the form of Tweets extracted from the website. The results summarize the impact of the initial months of the pandemic on the public sentiment, which can be useful for healthcare workers, authorities and researchers.

Introduction

The ongoing pandemic of novel coronavirus Disease, COVID-19 pandemic has left a devastating impact on the global economy, healthcare and many more sectors. With countries trying to keep up with the soaring cases, the implementation of lockdowns and containment measures was carried out. In the face of this rapidly changing situation, people all over the world felt different impacts on their personal lives and mental health. Recording these impacts is essential not only for solving the current problems, but also to document for creating robust support plans in the post pandemic world. The continuous analysis of trends in sentiment can help the authorities understand the impact of policies on the public sentiment as well.

To analyse the sentiment, we require continuous global data, either directly or indirectly from the people affected by the pandemic. In this study, we proposed to use the Twitter data in the form of tweets written in the English language from across the globe for our preliminary analysis, as Twitter is widely used by people across the world and contains sufficient textual content to conduct statistical analyses. We have employed a natural language processing pipeline that converts these tweets into interpretable insights for the authorities and healthcare workers. For this paper, we have focused on a subset of the Twitter data from the Indian subcontinent, to explore the trends in public sentiment. Our main contributions in the ongoing research include:

* Cleaning and preprocessing Twitter data to render it usable for statistical analysis.
* Topic modelling using Negative Matrix Factorization technique and Bag of Words model to find the topics commonly discussed on social media.
* Sentiment analysis using different machine learning algorithms and comparison of public sentiment in the 4 lockdowns in India.

In our future work we propose to:

* Leverage Hidden Markov models for POS tagging, and build network graphs to associate the popular topics with clusters of sentiment in different counties.
* Developing a model for Inferring causal inference in time series data to study the relationships between the number of cases and the public sentiment.

Related Work

Applying natural language processing to extract meaningful insights about the pandemic effects and policies has been an area of rising interest in the research community. Buckman et all [1] found a pronounced drop in consumer sentiment, by combining news sentiment and survey based consumer sentiment data. Alamoodi et all [2] discovered patterns in pandemic related literature, and used sentiment analysis to gain useful information. An interesting case study of the sentiment during lockdown in India found that the prominent sentiment was positive. [3]One work applied Sentiment Classification using Recurrent Neural Network, and found visible trends in positive and negative sentiment, along with occasional periods of stagnation where the public got bored with the pandemic.[4]

For analyzing the impact of the pandemic in India, using Twitter data has been a popular resource. For the purpose of gaining greater insight on the pandemic, Lamsal created an India-specific COVID-19 tweets dataset.[5] This dataset was used for multiple research applications including TClustVID, a novel model for analyzing COVID-19-related public tweets which found topics like “awareness”,”panic” and “coronavirus” in positive negative and neutral clusters. [6] Other works included analyzing the misinformation sources,[7] relating the COVID-19 cases with sentiment[8] and a temporal examination of the trending Twitter discussion on COVID-19. [9] We analyzed several papers on this subject and found interesting methodologies put to use for finding the impact of the pandemic globally.

Data Description

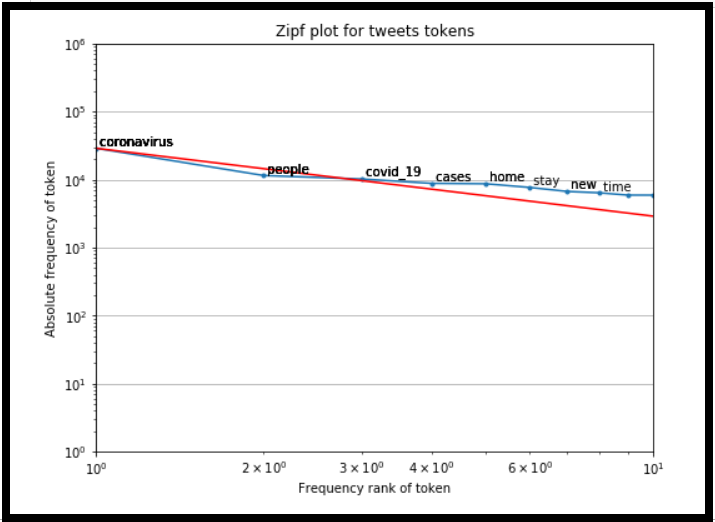
For this case study we combined data from 2 different sources for the sentiment and statistical analysis. The first dataset contains weets of users who have applied the following hashtags: #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid\_19, #epitwitter, #ihavecorona sourced from [10], using the rtweet package on CRAN. The second dataset was for the India specific analysis during the lockdowns, [5] from March 25, 2020 to May 23, 2020 scraped for over 100 hashtags including "corona", "#corona", "coronavirus", "#coronavirus", "covid". We combined these two datasets and obtained a dataset of 284,643 tweets across the globe, but mostly from India focussing on the 4 lockdown periods. The dataset was preprocessing extensively, including the stages of tokenizing, lemmatizing, removing stop words, removing hashtags, extracting topics, removing emojis, parsing paragraphs, correcting spelling errors, abbreviations and miscellaneous words.

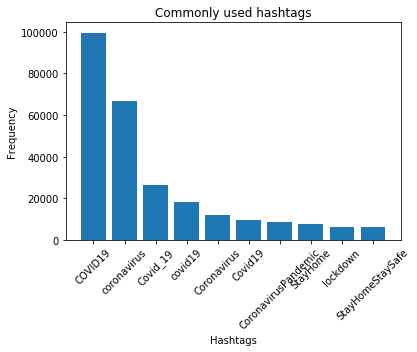
Methodology

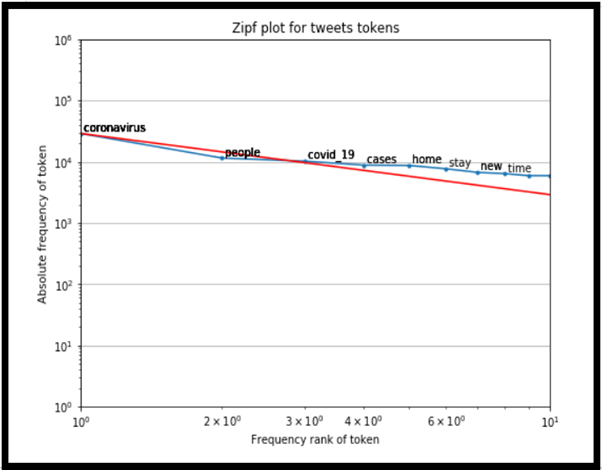
We propose the use of transformer based BERT models and classic supervised learning algorithms for sentiment analysis. For topic modelling we have leveraged the Non-Matrix Factorization and Bag of Words models to obtain the most important topics found in the tweets. As this is an ongoing research, in the future work, we will use Hidden Markov Models, causal inference algorithms and time series data models for the analysis.

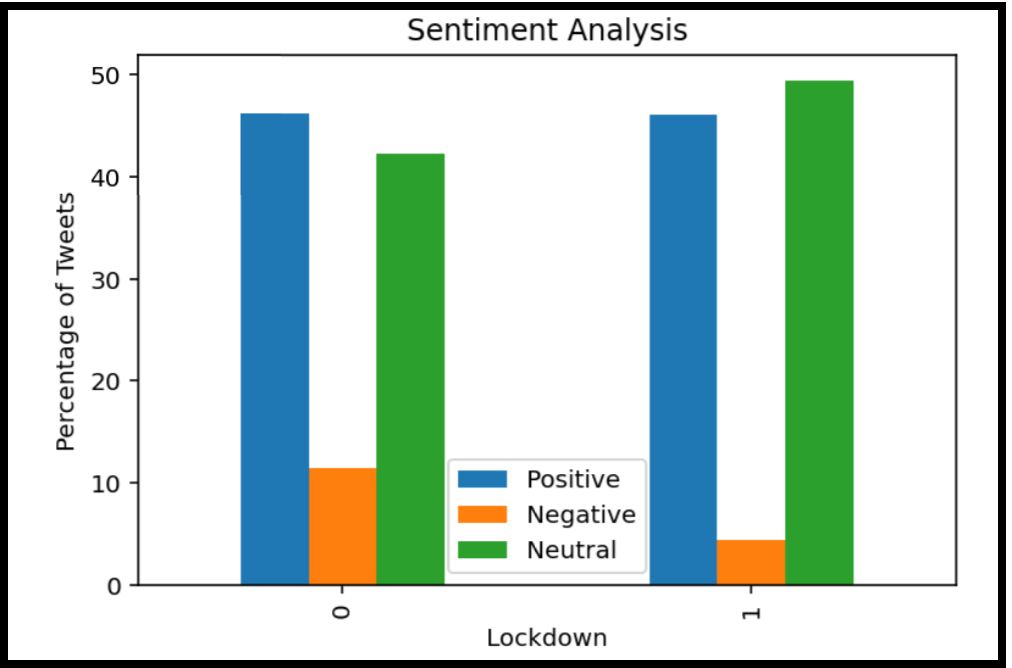
Results

This study provided us with some meaningful insights about the Twitter dataset, and the effect of the pandemic on the public sentiment. The NMF and Bag of words models provided the most commonly spoken about topics including: coronavirus,corona, people, covid, symptoms, virus and government. In India the topic modelling produced the terms like covid, delhi, maharashtra, hospital and liquor. The classification algorithms produced different results when tested against the ground truth. The BERT transformer-based model produced 67% accuracy while SVM classifier gave 80% accuracy.









Discussion and Future Work

We have presented results for topic modelling and sentiment analysis on Twitter data from across the world. We use a previously proposed state-of-the-art transformer model and a classic supervised learning algorithm, and compare their performance for sentiment analysis for the classification tasks: negative, positive and neutral. In future work, we will explore richer linguistic analysis, causal inference, building network graphs and finding the relationships between the sentiment trends and number of cases. We aim to convert the predictions in this paper into indicator inputs for more complex models, and to analyze the impact of policies like lockdowns in countries other than India as well. We hope that the purpose of this study to determine the nature of sentiment dynamics surrounding a pandemic, and to summarise the voice of the public will be very helpful for the authorities and researchers studying the effects of the pandemic.

References

[1]@article{buckman2020news,

title={News sentiment in the time of COVID-19},

author={Buckman, Shelby R and Shapiro, Adam Hale and Sudhof, Moritz and Wilson, Daniel J and others},

journal={FRBSF Economic Letter},

volume={8},

pages={1--05},

year={2020},

publisher={Federal Reserve Bank of San Francisco}

}

[2]@article{alamoodi2020sentiment,

title={Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review},

author={Alamoodi, Abdullah and Zaidan, Bilal and Zaidan, Aws and Albahri, Osamah and Mohammed, Khaled and Malik, Rami and Almahdi, Esam and Chyad, Mohammed and Tareq, Zaidoon and Albahri, Ahmed and others},

journal={Expert systems with applications},

pages={114155},

year={2020},

publisher={Elsevier}

}

[3]@article{barkur2020sentiment,

title={Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India},

author={Barkur, Gopalkrishna and Vibha, Giridhar B Kamath},

journal={Asian journal of psychiatry},

volume={51},

pages={102089},

year={2020},

publisher={Elsevier}

}

[4] @article{doi:10.1080/24751839.2020.1790793,

author = { László Nemes and Attila Kiss },

title = {Social media sentiment analysis based on COVID-19},

journal = {Journal of Information and Telecommunication},

volume = {5},

number = {1},

pages = {1-15},

year = {2021},

publisher = {Taylor & Francis},

doi = {10.1080/24751839.2020.1790793},

URL = { https://doi.org/10.1080/24751839.2020.1790793

},eprint = { https://doi.org/10.1080/24751839.2020.1790793}}

[5]@data{k8gw-xz18-20,

doi = {10.21227/k8gw-xz18},

url = {https://dx.doi.org/10.21227/k8gw-xz18},

author = {Rabindra Lamsal },

publisher = {IEEE Dataport},

title = {Tweets Originating from India During COVID-19 Lockdowns},

year = {2020} }

[6]@article {Satu2020.08.04.20167973,

author = {Satu, Md. Shahriare and Khan, Md. Imran and Mahmud, Mufti and Uddin, Shahadat and Summers, Matthew A. and Quinn, Julian M.W. and Moni, Mohammad Ali},

title = {TClustVID: A Novel Machine Learning Classification Model to Investigate Topics and Sentiment in COVID-19 Tweets},

elocation-id = {2020.08.04.20167973},

year = {2020},

doi = {10.1101/2020.08.04.20167973},

publisher = {Cold Spring Harbor Laboratory Press},

abstract = {COVID-19, caused by the SARS-Cov2, varies greatly in its severity but represent serious respiratory symptoms with vascular and other complications, particularly in older adults. The disease can be spread by both symptomatic and asymptomatic infected individuals, and remains uncertainty over key aspects of its infectivity, no effective remedy yet exists and this disease causes severe economic effects globally. For these reasons, COVID-19 is the subject of intense and widespread discussion on social media platforms including Facebook and Twitter. These public forums substantially impact on public opinions in some cases and exacerbate widespread panic and misinformation spread during the crisis. Thus, this work aimed to design an intelligent clustering-based classification and topics extracting model (named TClustVID) that analyze COVID-19-related public tweets to extract significant sentiments with high accuracy. We gathered COVID-19 Twitter datasets from the IEEE Dataport repository and employed a range of data preprocessing methods to clean the raw data, then applied tokenization and produced a word-to-index dictionary. Thereafter, different classifications were employed to Twitter datasets which enabled exploration of the performance of traditional and TclustVID classification methods. TClustVID showed higher performance compared to the traditional classifiers determined by clustering criteria. Finally, we extracted significant topic clusters from TClustVID, split them into positive, neutral and negative clusters and implemented latent dirichlet allocation for extraction of popular COVID-19 topics. This approach identified common prevailing public opinions and concerns related to COVID-19, as well as attitudes to infection prevention strategies held by people from different countries concerning the current pandemic situation.Competing Interest StatementThe authors have declared no competing interest.Funding StatementNo external funding was receivedAuthor DeclarationsI confirm all relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained.YesThe details of the IRB/oversight body that provided approval or exemption for the research described are given below:All relevant ethical guidelines have been followedAll necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.YesI understand that all clinical trials and any other prospective interventional studies must be registered with an ICMJE-approved registry, such as ClinicalTrials.gov. I confirm that any such study reported in the manuscript has been registered and the trial registration ID is provided (note: if posting a prospective study registered retrospectively, please provide a statement in the trial ID field explaining why the study was not registered in advance).YesI have followed all appropriate research reporting guidelines and uploaded the relevant EQUATOR Network research reporting checklist(s) and other pertinent material as supplementary files, if applicable.YesExperimental data has been gathered from IEEE Dataport},

URL = {https://www.medrxiv.org/content/early/2020/08/04/2020.08.04.20167973},

eprint = {https://www.medrxiv.org/content/early/2020/08/04/2020.08.04.20167973.full.pdf},

journal = {medRxiv}

}

[7]@article{silva2020predicting,

title={Predicting Misinformation and Engagement in COVID-19 Twitter Discourse in the First Months of the Outbreak},

author={Silva, Mirela and Ceschin, Fabr{\'\i}cio and Shrestha, Prakash and Brant, Christopher and Fernandes, Juliana and Silva, Catia S and Gr{\'e}gio, Andr{\'e} and Oliveira, Daniela and Giovanini, Luiz},

journal={arXiv preprint arXiv:2012.02164},

year={2020}

}

[8]@article{luu2020relationship,

title={The relationship between sentiment score and COVID-19 cases in the USA},

author={Luu, Truong Jack P and Follmann, Rosangela},

year={2020}

}

[9]@article{chandrasekaran2020twitter,

title={Twitter talk on COVID-19: a temporal examination of topics, trends and sentiments},

author={Chandrasekaran, Ranganathan and Mehta, Vikalp and Valkunde, Tejali and Moustakas, Evangelos},

journal={Journal of Medical Internet Research},

year={2020},

publisher={JMIR Publications}

}

[10]<https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april>