electronics

MDPI

Article

Measuring the Impact of Language Models in Sentiment Analysis for Mexico's COVID-19 Pandemic

Edgar León-Sandoval, Mahdi Zareei, Liliana Ibeth Barbosa-Santillán * and Luis Eduardo Falcón Morales

School of Engineering and Sciences, Tecnologico de Monterrey, Zapopan 45201, Mexico
* Correspondence: ibarbosa@tec.mx

Abstract: The world has been facing the COVID-19 pandemic, which has come with an unprecedented impact on general physical health and financial and social repercussions. The adopted mitigation measures also present significant challenges to the population's mental health and health-related programs. It is complex for public organizations to measure the population's mental health to incorporate its feedback into their decision-making process. A significant portion of the population has turned to social media to express the details of their daily life, making these public data a rich field for understanding emotional and mental well-being. To this end, by using open sentiment analysis tools, we analyzed 760,064,879 public domain tweets collected from a public access repository to examine the collective shifts in the general mood about the pandemic evolution, news cycles, and governmental policies. Several modern language models were evaluated and compared using intrinsic and extrinsic tasks, that is, the sentiment analysis evaluation of public domain tweets related to the COVID-19 pandemic in Mexico. This study provides a fair evaluation of state-of-the-art language models, such as BERT and VADER, showcasing their metrics and comparing their performance against a real-world task. Results show the importance of selecting the correct language model for large projects such as this one, for there is a need to balance costs with the model's performance.

Keywords: sentiment analysis; language model evaluation; big data; COVID-19; machine learning; Mexico; twitter

Citation: León-Sandoval, E.; Zareei, M.; Barbosa-Santillán, L.I.; Falcón Morales, L.E. Measuring the Impact of Language Models in Sentiment Analysis for Mexico's COVID-19 Pandemic. *Electronics* **2022**, *11*, 2483. https://doi.org/10.3390/electronics11162483

Academic Editors: Diego Reforgiato Recupero, Danilo Dessi' and Harald Sack

Received: 24 June 2022
Accepted: 28 July 2022
Published: 10 August 2022

1. Introduction

As the world has been facing the ongoing COVID-19 (COronaVIrus Disease 2019) pandemic, governments and public and private organizations must prioritize public well-being in their decision-making process. The challenges the COVID-19 pandemic presents, both in individuals' emotional and psychological well-being, raise the need to collect health-related data and, moreover, to build dashboards that show critical information, make it easily accessible, and gather the day-to-day data of the pandemic progression and ongoing infection rates and fatality, among other statistics. However, emotional health, previously studied in other geographic locations such as the United States [1] and Mexico [2] has also shown to have long-term implications for the well-being of the populations and how it is affected by news and government decisions related to the pandemic. This situation presents the need for a tool to measure the impact of communications transmitted to the population, which can also serve as a feedback mechanism to better adjust future announcements. Such a mechanism faces many challenges: data recollection, processing capabilities, and deciding what measurement instrument to utilize.

Acquiring and processing this amount of information is not easy, as this is a perfect example of the challenges encountered by the three Vs of big data: volume, variety, and velocity [3]. We adhere to the most common definition of big data based on *the three Vs*, first introduced by [4]. However, there are multiple definitions containing different aspects of these architectures, such as analysis, value, computer power, visualization, variability,

and veracity, among various others. An in-depth description of these definitions is described by [5], but suffice to say that this research work adheres to the big data concept by most definitions. It presents several problems, such as those listed next:

- Acquiring feedback on the emotional state is both expensive and time-consuming.
- Having these data available presents a significant challenge in processing capabilities and comes with a long wait time to receive feedback.
- Building such systems is costly, regardless of the volume variations in the data.

Traditional survey methods, such as interviews or surveys, are prohibitive. Besides the high expense, they require significant time to gather feedback on a small portion of the population, providing information on discrete periods rather than a continuous flow. Twitter is a mature, well-established, and popular microblogging service that offers users a platform to share their opinions, conversations, reviews, and other information. A large corpus of heterogeneous data was collected [6], which we refer to as the COVID-19 Twitter chatter dataset. It includes raw text, tweet metadata, images, videos, URLs, and popularity. This corpus is an excellent candidate for performing sentiment analysis to follow public opinion on any given topic or event but presents several challenges, including the high computing resources needed for the research and a curated, well-defined training corpus. Furthermore, the advances in technology nowadays allow the processing of data in large volumes, at a fast velocity, and from numerous heterogeneous sources, making possible the analysis of sentiments on a near real-time basis [7].

Sentiment analysis is a discipline that allows the determination of the sentiment classification and polarity of any given free-form text. While there are multiple sentiment analyzers, they all share the same basic pipeline: preprocessing, where we normalize the input text and transform it into a form that a machine can process; the actual sentiment classification or polarity determination, which a language model calculates; and postprocessing, where the outputs need further normalization or interpretation. We often find two types of language models: rule-based models, where we try to construct an expert system to interpret the text, or statistical-based models, often used with deep learning architectures, which provide the most robust and best-performing models to date. However, we train and evaluate these models using a curated corpus of data, from which we calculate their intrinsic measurements. These intrinsic measurements can be optimistic and do not guarantee that a well-performing model will remain so when used on real-world data, for the model will evaluate data significantly different from the data used during training.

For these reasons, we provide an empirical study comparing these language models to each other in terms of this real-world extrinsic task: the emotional response of the Mexican population by performing a sentiment analysis on COVID-19-related tweets. In addition to presenting a summary of each model's measurements, we compare how this affects the said study, highlighting their relevance and cost/performance trade-offs. This work is organized as follows. The introduction showed the context of the problem, a brief overview of the proposed solution, a general background of existing COVID-19 studies based on Twitter data, and an overview of the language models to be evaluated. The Methods section highlights the evaluation process and an in-depth review of the language models included in this work. Experiment and Results show the measurements taken and the results gathered, with some explanations of the meanings and interpretations drawn from the results. Finally, we close this work with a brief discussion of the conclusions drawn here, as well as future steps that can be taken to explore this area further.

Literature Review

Next, a quick literature review is presented, dividing the literature into two parts: language model evaluation and sentiment analysis performed for COVID-19-related data taken from Twitter. Natural language processing has recently seen novel architectures implement language models in a way that provides robustness and accuracy. All implementations use nonlinear statistical models as language representations, in different ways, from vast attention-based deep learning architectures to simpler dictionary-based deploy-

ments, such as VADER [8]. VADER is an open-source, rule-based robust language model that can handle commonly employed complex grammar structures commonly used in social media. For training data, it utilizes a curated corpus evaluated by humans. BERTweet [9] is based on BERT [10], using a pretraining procedure somewhat similar to that utilized by RoBERTa [11], both of which use publicly available tweets in English for training and evaluation. TimeLMs [12] introduces a time concept into the language model by utilizing continuous learning and thus accounts for future and out-of-distribution tweets it might encounter. This language model also uses publicly available tweets in English for training and evaluation.

The implementer of each novel language model architecture provides a set of metrics that serve as the basis for each of the improvements provided. These are intrinsic metrics, for they evaluate the model's performance against a previously defined test corpus, which is usually part or a superset of the training dataset. Table 1 provides a summary of these metrics for different implementations of language models. However, the behavior of the models can be different when faced with the real world, as the data might present another distribution or present new cases that were not part of the training. This makes it indispensable to evaluate models using real-world data on real-world conditions with a real task. This evaluation method is known as an extrinsic metric, where the model is not directly evaluated but compared with other models based on its performance on a real-world task. For example, TweetEval [13] evaluates a few selected language models, with the single exception of VADER which was measured in a different study [14], training and evaluating them under a curated corpus appropriate for the task [15], collecting metrics, and concluding on a single clear winner. However, we have no information on how such a winner model would behave with accurate, non-curated data.

Table 1. Summary of different popular language models evaluated using a sentiment analysis corpus [15]. The evaluation of VADER was performed [14].

Author	Model Name	Summary	Score
Nguyen et al. [9]	BERTweet	Uses pretraining over BERT for Tweets in English	73.4
Barbieri et al. [13]	RoBERTa on Twitter	A RoBERTa variant trained on a Twitter corpus	69.1
Hutto and Gilbert [8]	VADER	Uses rule-based evaluation and human-tagged data	69

Table 2 displays a summary of sentiment analysis studies performed on Twitter data related to COVID-19, where we found them to utilize a small dataset, either in volume or in the length of the analyzed time frame. In general, only tweets written in English were accepted, restricted to the U.S., with just a few exceptions. For example, the study in [16] restricts the studied tweets to those that originated only in Australia, in contrast to the survey [17], which uses a global dataset. Note that most of these studies [1,16–21] focus on panel data analysis, except for [2,22–26], which use a time series analysis. There have been other works that provide additional information; for example, the study [27] in the U.S. and [28] in Canada, which provide ample evidence on the correlation of beliefs shared on Twitter and the social distancing practices in real life, providing good indicators for risk management. Since the study [2] also provides a public dataset and enough details of the technology and language model utilized, we focused our methodology on matching it as closely as possible and performing comparisons with the other models accordingly.

Table 2. Summary of studies evaluating sentiment polarity over COVID-19-related tweets.

Author	Summary	Dataset		
		Location	Time Span (in 2020)	*n*
Adikari et al. [16]	Topic analysis, follows popular subjects, pre/post lockdown	Australia	Jan to Sep	73 K
Abd-Alrazaq et al. [18]	Uses PostgreSQL and topic analysis, pre/postlockdown		Feb 2 to Mar 15	167 K
Boon-Itt and Skunkan [19]	Topic analysis, 3 panel data analysis	US	Dec 13 to Mar 9	108 K
Lwin et al. [17]	Uses the Plutchik basic sentiments, pre/postlockdown	Global	Jan 28 to Apr 9	20 M
Xue et al. [20]	Topic analysis, pre/postlockdown	US	Mar 7 to Apr 21	4 M
Valdez et al. [21]	Topic analysis, pre/postlockdown, follows popular subjects	US	Jan 28 to Apr 7	86 M
Huerta et al. [1]	Pre/postlockdown	MA, US	Jan 1 to May 14	2.88 M
Crocamo et al. [22]	Time series		Jan 19 to Mar 2	3.3 M
Chandra and Krishna [23]	Time series, topic analysis	India	Mar to Sep	150 K
Alam et al. [24]	Time series, topic analysis		Dec 21 to Jul 21	125.9 K
Garcia and Berton [25]	Time series, topic analysis, multilanguage	US and Brazil	Apr to Aug, 2021	6.5 M
Singh et al. [26]	Time series, topic analysis		May 3 to Aug 29, 2021	400 K
León-Sandoval [2]	Time series, multilanguage	Mexico	Feb 1 to Dec 31	760 M

2. Data and Methods

We collected a dataset of tweets from an open-access repository of global COVID-19-related tweets, which we refer to as the COVID-19 Twitter chatter dataset [6]. The dataset was designed to collect every tweet related to the COVID-19 pandemic and included metadata to facilitate analysis and filtering before consumption. This COVID-19 Twitter chatter dataset provides a broad collection of tweet IDs, geographical locations, and detected language, utilizing the following schema: *[tweet_id, date, time, lang, country_code]*. However, we encountered schema inconsistencies over time. For example, the annotation of *country_code*, which is necessary for filtering before requesting a tweet lookup, was not introduced until the second half of the year, and even then, a large number of tweets lacked this metadata annotation.

For this reason, we had to load them via Twitter's public API to filter out tweets originating from outside Mexico, which may have left out data from those users who chose not to share their location. We used this information to download each tweet in Mexico, discarding all other metadata provided by Twitter's API for privacy reasons. Specifically, we retrieved COVID-19-related tweets posted in Mexico from 1 February 2020 to 31 December 2020. All tweets were scrubbed of personally identifiable information to ensure user privacy and comply with ethical practices in social networks, resulting in the following simplified schema: *full_text, id, time_stamp*. It is worth mentioning that this dataset included tweets in both English and Spanish, for a large part of the population engages on social media in English. This methodology followed the same consumption strategy previously followed [2].

Figure 1 shows the data ingestion pipeline, for which we used the regular lookup V2 API. Note that the resulting sample size for the dataset was quite large, consisting of n = 2,142,800 unique tweets, resulting in an ample sample to perform this analysis. Previous studies, summarized in Table 2, used large-scale sentiment analysis to accurately predict public mood and how it applied to several domains, including those of emotional and psychological well-being [29].

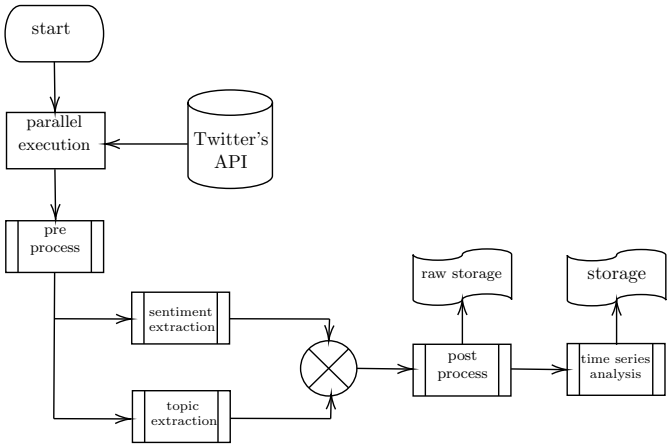

Figure 1. *Data Flow Overview.* We process the data in three main stages: first, we load the desired tweet IDs from the COVID-19 Twitter chatter dataset then consult them directly from Twitter using the official APIs. For pre-processing, we clean up and filter the data, and then we process this dataset to produce a time series of the perceived COVID-19-related sentiment.

Natural language processing has recently seen novel architectures implement language models, increasing robustness and accuracy for several natural language tasks, such as sentiment polarity determination. These implementations, particularly those compared in this work, are listed in Table 1. All implementations use nonlinear statistical models as language representations, in different ways, from vast attention-based deep learning architectures to more straightforward dictionary-based implementations, such as VADER. These are the language models evaluated in this work, and we briefly describe each. VADER [8] is an open-source rule-based robust language model that can handle complex grammar structures commonly employed in social networks. VADER is reliable, fast to deploy, and needs few resources to evaluate new text entries. However, for training, it utilizes a curated corpus evaluated by humans, making adapting it, or incorporating new data, a difficult task. BERTweet [9] is based on BERT [10], using a pretraining procedure similar to that utilized by RoBERTa [11] and uses publicly available Tweets in English for training and evaluation. TweetEval [13] already has scored and compared both BERTweet and RoBERTa, finding better performance in this particular task in the former. Both provide a robust language model, which is enormous both in size and evaluation resources needed. Still, we can quickly update the model if required by exploiting pretraining and multilanguage support. TimeLMs [12] introduces a time concept into the language model by utilizing continuous learning, gaining the ability to account for both future and out-of-distribution tweets the model might encounter. The TimeLMs language model also uses publicly available tweets in English for training and evaluation. This results in a robust language model that prefers novel entries and can deal with out-of-distribution evaluations. However, it is susceptible to adversary attacks, and its performance can fall over time. Although we can mitigate this issue by following standard MLops practices, this last language model is not part of this study.

$$F1\_score = 2 * \frac{precision * recall}{precision + recall} \tag{1}$$

TweetEval [13] proposes a metric comparing multiple language models with each other, evaluated using a properly curated corpus provided by SemEval [15], from which we obtained the intrinsic measurements for all models except VADER, for which [14] calculated its measurement. This strategy is appropriate as TweetEval uses a standardized protocol consisting of seven NLP tasks, one being sentiment analysis, and uses Twitter corpora to train the models. It also provides a single global metric, which is handy though somewhat straightforward, as it averages the scores for each task [30]. The scoring used standard

averaged F1 scores, an harmonic mean of precision and recall defined by Equation (1), for most tasks except sentiment analysis, which relies on recall alone. The results for the sentiment analysis task are in Table 1, along with a small summary of the models. However, while there is extensive use of Twitter corpora in training these language models, the measurements are still considered to be intrinsic, which leaves the question of the performance of these models over real-world data open. This remains true even if the data in question are another Twitter corpora.

For the time series analysis, we employed a similar methodology to that of [2], presented next. We consumed data using the Twitter public API, which were then stored in hard storage, triggering change events that fed the entries into a data pipeline, making it easy to swap them for near-real-time tweet streams. Data were then cleaned and stored in a sizeable non-SQL database, from which we queried data for exploration, experimentation, or model training. Once again, we used data triggers to feed them into the sentiment polarity calculator, ending in another large non-SQL database instance. This final instance was a source for aggregation and analysis, from which we could calculate daily aggregates. For the time series analysis, we started by denoising the series. We opted for a moving average of seven days as we also found solid weekly seasonality in the data. Next, the data were detrended by fitting a regular time series model. Several partial autocorrelation tests were performed to find a good initial parameter approximation and validate the model. The residuals and the box tests revealed a good fit of the model, resulting in a $p\text{-}value < 2.2 \times 10^{-16}$ for the sentiment polarities calculated by all models. These steps were repeated for several aggregation statistics, keeping the mean and the standard deviation relevant, as they summarized the behavior observed in the data well. More details are available in the *Experiments* section.

This study also followed the technical implementation suggested by [2], as it is easy to replicate and flexible enough to alter without significant changes for our purposes. The system was implemented on top of Google cloud services (GCP), allowing a tight `TensorFlow` integration, loose coupling, and dynamic scaling, and was written in `Python 3.6`, with its data-focused libraries, such as `TensorFlow`. Using this technology allowed the integration of MLops practices, making it easy to update models. Figure 2 shows an overview of this technical architecture. The general flow is as follows:

1. Data are ingested directly from Twitter, using the identifiers provided by the *COVID-19 Twitter chatter dataset* and the general query API provided by Twitter.
2. We then publish the tweet in `Pub/Sub`, which is written directly in cloud storage for future reference and debugging.
3. `Pub/Sub` feeds this data entry into a serverless function, which then evaluates the tweet polarity using: VADER [8], BERTweet [9], and RoBERTa [11] implementations, all of them written in Python and `TensorFlow`, and post the results again in `Pub/Sub` to be fed into `BigTable` for final consumption.
4. The data are now ready for consumption by a managed `Dataproc` instance with two different approaches:
   (a) One where we execute periodic batch jobs to calculate the daily aggregations, stored in cloud storage and `BigTable` for easy access.
   (b) Another is where we publish dynamic and executable Jupyter notebooks for manual data exploration.

Regarding the time series, VADER uses a composed metric obtained by normalizing both the positive and negative scores and using an external factor to better approximate a 1 to -1 distribution [8],

$$norm\_score = \frac{sum\_polarities}{\sqrt{sum\_polarities^2 + \alpha}} \quad (2)$$

where $sum\_polarities$ is the simple addition of positive and negative polarities, and α is initialized as $\alpha = 15$. We need to adjust this α for every operation based on a heuristic

and the lexicon collected by the language model. This alone gives us a clue regarding the differences found in the time series, for the deep-neural-network-based language models implement a $softmax$ layer that already provides this normalization for us. However, we considered this difference by adjusting the results accordingly, calculating the α used by VADER to replicate its distribution approximation used for normalization. Although this would not be necessary for either BERTweet or RoBERTa, as the deep learning network uses a $softmax$ layer to calculate this distribution, we wanted to match the methodology as much as possible. Still, it is something to keep in mind moving forward.

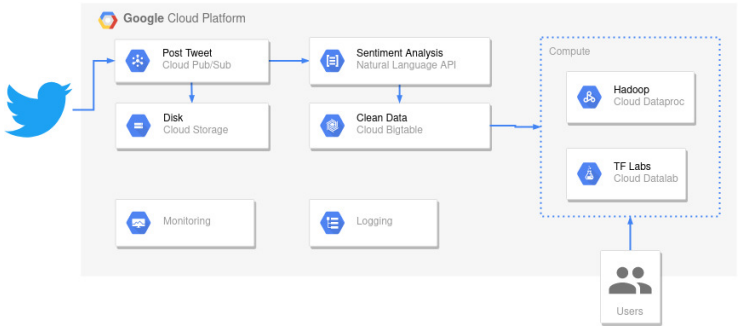

Figure 2. *General Architecture implemented in Google Cloud Platform (GCP)* [2]. This cloud-based architecture ingests tweets using the official Twitter APIs, sending each one of these through Google's `Pub/Sub`, which uses as endpoints essential preprocessing, raw storage, and a serverless function to calculate the sentiment polarity. This function's results are fed into `BigTable` through another `Pub/Sub` pipeline.

With these adjustments in place, we generated the time series using a rolling average of 7 days to denoise the data, which we found to work best, and it was aligned with the seasonality found [31] while also matching the strong weekly seasonality found in this analysis. Then, we detrended the data by fitting a regular time series model. We performed several partial autocorrelation tests to find a good initial parameter approximation and validate the ARIMA model, revealing good results using the residuals, providing seasonality and trends. Although there was a precise offset in the data, the model fit, trends, and seasonality followed by the time series were similar. We confirmed this by conducting a simple Pearson correlation test, which yielded low correlation levels on the time series with the lag present, which was not an issue in our case. We present more details in the Experiments and Results sections.

3. Experiments

We performed a sentiment analysis on COVID-19-related tweets posted in Mexico from 1 February 2020 to 31 December 2020, forming a corpus of 760,064,879 tweets, which after preprocessing and filtering came to a total of n = 2,142,890 utilized tweets, retrieved from the *COVID-19 Twitter chatter dataset* [6]. Note that the ranges of polarity values went from -1 (i.e., entirely negative) to 1 (i.e., completely positive), where 0 was considered a neutral value or an utterly objective tweet (but given that these tweets were for the most part opinions, this was rarely the case). We performed this sentiment polarity determination using three language models: VADER, BERTweet, and RoBERTa. Table 3 presents a monthly summary of the sentiment polarity for a given month, but the analysis was performed with daily granularity.

Table 3. Monthly summary statistics for the compound sentiment polarity in Mexico, utilizing several language models. Note that the polarity is expressed in the range $(-1, 1)$.

Month	VADER			BERTweet			RoBERTa		
	μ	p_{50}	σ^2	μ	p_{50}	σ^2	μ	p_{50}	σ^2
February	−0.0984257	−0.0429283	0.0751305	−0.161653	−0.279797	0.166135	−0.329798	−0.331657	0.0751666
March	−0.180352	−0.296000	0.171638	−0.0670456	−0.0105233	0.0825488	−0.248890	−0.237242	0.0988796
April	−0.175834	−0.286133	0.186721	−0.0405222	0.00839683	0.0837904	−0.197245	−0.166518	0.108357
May	−0.135060	−0.266584	0.227397	0.0414142	0.0569896	0.0990766	−0.120022	−0.121882	0.171598
June	−0.187039	−0.302487	0.214230	0.0349433	0.0514157	0.0933161	−0.117042	−0.106476	0.165380
July	−0.171287	−0.296000	0.216006	0.0297321	0.0500704	0.0890589	−0.106039	−0.0941371	0.155880
August	−0.156576	−0.283995	0.220971	0.0338199	0.0504739	0.0920116	−0.105613	−0.0978736	0.158233
September	−0.137832	−0.284002	0.231412	0.0276182	0.0598745	0.112541	−0.101315	−0.113173	0.179672
October	−0.144219	−0.277552	0.223694	0.0351052	0.0524153	0.0972368	−0.0956743	−0.0840913	0.165041
November	−0.157551	−0.283380	0.216210	0.0507385	0.0620046	0.0995041	−0.0828981	−0.0851012	0.167781
December	−0.133056	−0.281765	0.228512	0.0493567	0.0687692	0.0951525	−0.0654149	−0.0741181	0.166066

We calculated a smoothed time series, where the box-test showed a $p\text{-}value < 2.2 \times 10^{-16}$ for all models, indicating a high probability of encountering autocorrelations in the data. This led to a further exploration using a partial ACF (auto correlation function). These ACF values are reported in the Appendix A. Here, we observed strong indications of weekly autocorrelations, which helped us quickly find the correct coefficients for fitting an ARIMA (autoregressive integrated moving average) model and decomposing the time series. Although the coefficients suggested a substantial similarity in the ARIMA models, there were some differences as well, more noticeable in the VADER model. Thus, we performed a Pearson correlation test for the different language models. These models should present no time lag nor noise in the time dimension, making this test a good candidate as opposed to a more time-consuming and resource-intensive test, such as the dynamic time warping distance [32].

Table 4 summarizes the Pearson correlation index between the language models for the compound sentiment polarity, while Table 5 shows the same correlations for the positive polarity only. This shows a strong correlation between BERTweet and RoBERTa, an expected behavior given that both are based on the BERT architecture and use Twitter data for fine-tuning. However, there is a stronger correlation with VADER for the positive polarity. Remember that compound polarity is calculated based on positive and negative polarities, and VADER uses a heuristic-based approximation replicated for both BERTweet and RoBERTa. We executed these experiments in the GCP pipeline described by Figure 2 using a CPU-only solution for VADER and a GPU configuration for evaluating both BERTweet and RoBERTa. Nevertheless, the time series analysis was evaluated in a CPU-only managed instance inside GCP. A more in-depth analysis and visualizations are presented in the next Section *Results*.

Table 4. Pearson's correlation coefficients for the compound sentiment polarity. Note the low correlation coefficient when comparing both BERTweet and RoBERTa language models with VADER.

	VADER	BERTweet	RoBERTa
VADER	1	0.3305117	0.3018875
BERTweet	0.3305117	1	0.9337619
RoBERTa	0.3018875	0.9337619	1

Table 5. Pearson's correlation coefficients for the positive sentiment polarity. Note the high correlations between all of the language models.

	VADER	BERTweet	RoBERTa
VADER	1	0.8715221	0.8809533
BERTweet	0.8715221	1	0.9763189
RoBERTa	0.8809533	0.9763189	1

4. Results

We performed a sentiment analysis on a corpus of 760,064,879 tweets posted from Mexico from 1 February 2020, to 31 December 2020 in multiple languages, which after

preprocessing and filtering came to a total of n = 2,142,890 utilized tweets, retrieved from the COVID-19 Twitter chatter dataset [6]. We performed this sentiment polarity determination using three language models, VADER, BERTweet, and RoBERTa. The resulting smoothed time series showed a $p\text{-}value < 2.2 \times 10^{-16}$, indicating a high probability of encountering autocorrelations in the data. Here, we observed strong indications of weekly autocorrelations, which helped us quickly find the correct coefficients for fitting an ARIMA (autoregressive integrated moving average) model and decomposing the time series. For comparison, we performed the Pearson correlation test, for we did not expect to see any time warping in the time series.

From the Figures 3–5, we observe a somewhat similar shape in the data, and all language models agree on the spikes caused by important events or governmental decisions, as reported [2]. For a detailed list of these events, please refer to [2], but they were all nationwide events or official announcements regarding COVID-19 by the relevant authorities. While the magnitude is not precisely the same, the spikes exist on similar dates, indicating a "good-enough" sensibility of the measurement tool to repeat the analysis and reach the same conclusions, at least as far as the sensibility of the data to important real-world events. Figure 6 presents the time series decomposition, that is, the raw data, the trend, seasonality, and Gaussian noise, from which we can observe a robust weekly seasonality and a trend of $m = 0.0004020335$. The appendix includes the same charts for VADER (Figure A1) and RoBERTa (Figure A2), excluded here for clarity, showing similar trends in all of the language models ($m = 0.00001110643$ for VADER, $m = 0.0004020335$ for BERTweet, and $m = 0.0006753483$ for RoBERTa) as well as a similar seasonality.

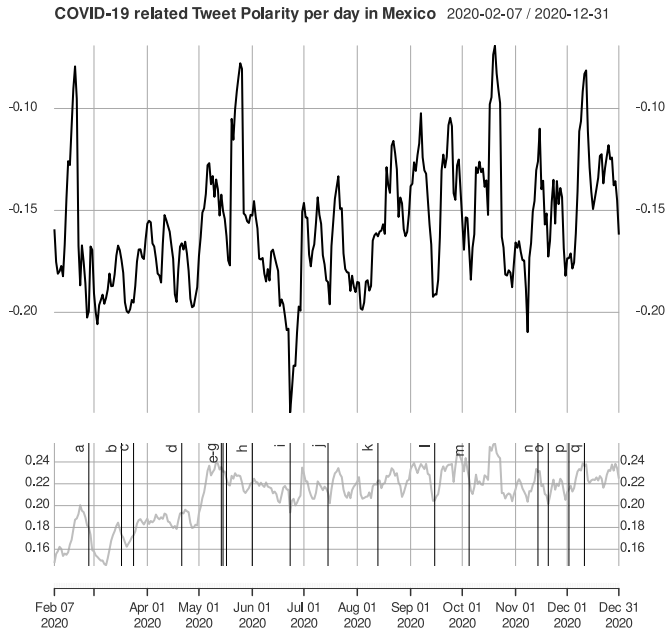

Figure 3. *Year–Long time series of daily averaged VADER's compound sentiment polarity of COVID-19-related tweets, detrended, with its variance and significant events in Mexico*. This time series is based on the same dataset collected [6] and restricted to Mexico from 2 February to 31 December 2020. Data were smoothed over via a 7-day rolling mean. Variance is also included for readability, as well as significant events found to cause an impact [2].

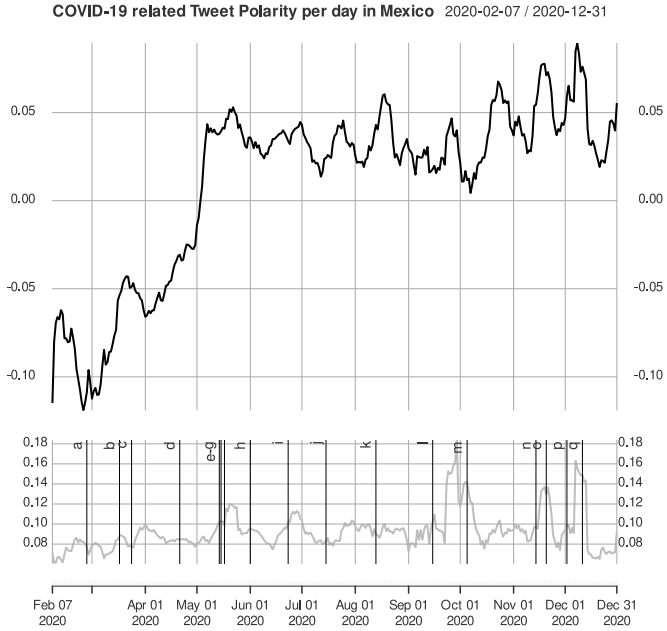

Figure 4. *Year–Long time series of daily averaged of BERTweet's compound sentiment polarity of COVID-19-related tweets, detrended, with its variance and significant events in Mexico*. This time series is based on the same dataset collected [6] and restricted to Mexico from 2 February to 31 December 2020. Data were smoothed over via a 7-day rolling mean. Variance is also included for readability, as well as significant events found to cause an impact [2].

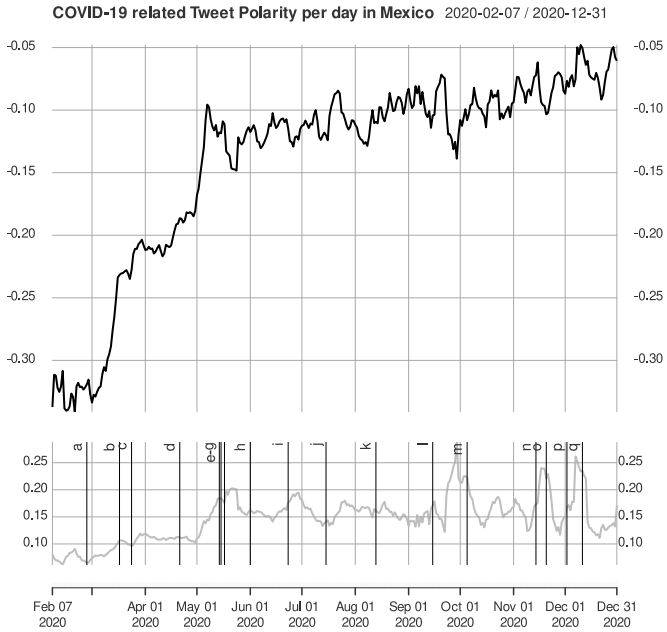

Figure 5. *Year–Long time series of daily averaged RoBERTa's compound sentiment polarity of COVID-19-related tweets, detrended, with its variance and significant events in Mexico*. This time series is based on the same dataset collected [6] and restricted to Mexico from 2 February to 31 December 2020. Data were smoothed over via a 7-day rolling means. Variance is also included for readability, as well as significant events found to cause an impact [2].

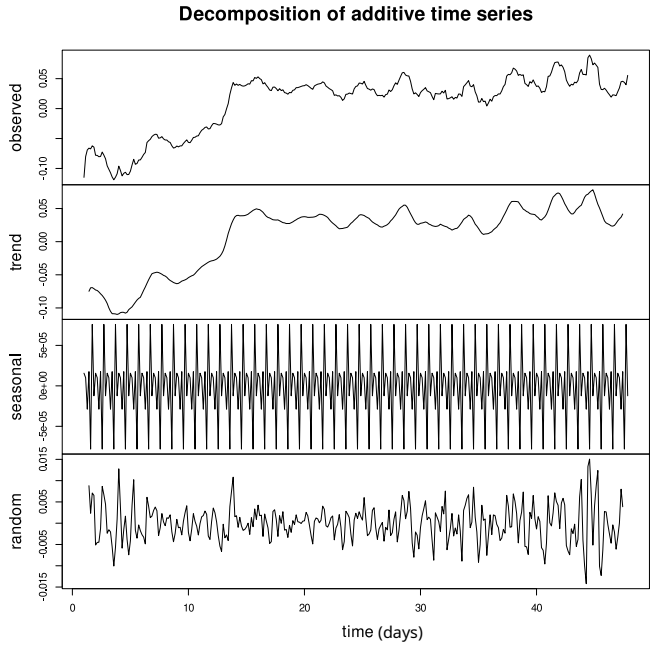

Figure 6. *Year–Long time series of daily averaged compounded sentiment polarity of COVID-19-related tweets, utilizing BERTweet in Mexico*. Shown is the decomposed sentiment analysis: the averaged time series, trends, weekly seasonality, and random data. The time axis is shown in days, starting from 2 February up to 31 December 2020.

Table 1 summarizes a compound metric evaluating different implementations of language models. This single global metric follows a similar methodology as GLUE [30] in that it uses an average of the relevant metrics for the given dataset. The metric utilized for this sentiment analysis task was the macroaveraged recall.

Figure 7 shows the daily average of the detrended compound sentiment polarity ranging from 2 February to 31 December 2020. Note that the Appendix A contains the average positive, negative, and raw compound sentiment polarity for the same period. This time series was smoothed over a 7-day rolling average, for we found a solid weekly seasonality, which matched the seasonality previously found [31] as well. We see an offset in each language model from here, but they follow a similar trend and present peaks at similar points, though their magnitude can be somewhat different, resulting in a similar sentiment intensity between VADER and RoBERTa, BERT having the group's least intense magnitude. This behavior was expected as the RoBERTa implementation uses English tweets as a dataset for fine-tuning. Tables 4 and 5 summarize the *Pearson's correlation test* indexes obtained when comparing the models. It is interesting to see a strong correlation with positive sentiments, particularly with RoBERTa, which should not be surprising as this model was also trained on Twitter data. However, when we calculated the compound sentiment polarity using VADER's coefficient (see Equation (2)), we saw low results typical of noisy signals. This would suggest that the most significant difference and drawback in utilizing VADER is its approximation function, defined in Equation (2), for calculating the compound sentiment polarity.

BERTweet and RoBERTa are large, deep-learning-based language models, sharing the same architecture BERT uses. A pretraining dataset of 80 GB is used for both, containing 850M English-only tweets. This architecture comes with two optimization levels, where BERTweet requires 135 M parameters and RoBERTa 355 M parameters, both for forward and backward propagation, resulting in an expensive training and evaluation architecture. Meanwhile, VADER needs human intervention to tag the data, but the model is a lightweight, rule-based language model. This makes VADER much cheaper to use and

deploy; however, as Figure 8 shows, the variance in BERT is significantly better than all other models, leaving VADER as the least stable model.

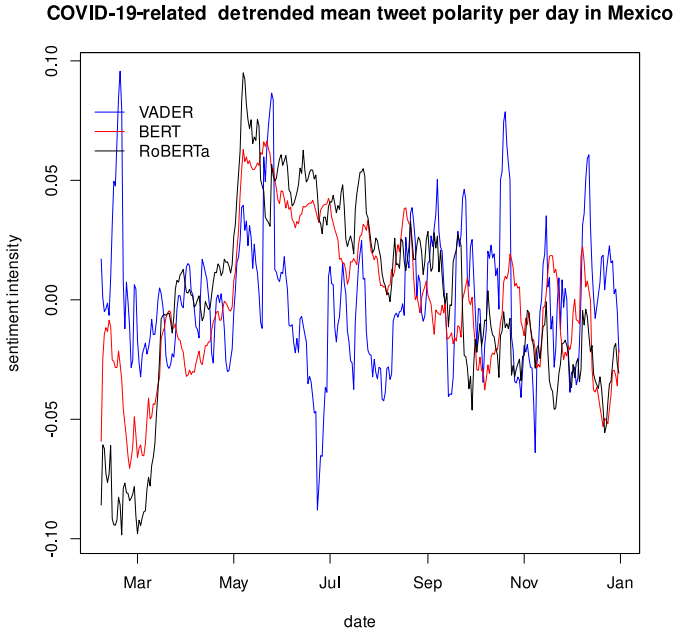

Figure 7. *Year–Long time series of daily averaged compound sentiment polarity of COVID-19-related tweets, detrended, in Mexico*. Each line represents a time series based on the same dataset collected [6] and restricted to Mexico, from 2 February to 31 December 2020. Data were smoothed via a 7-day rolling mean.

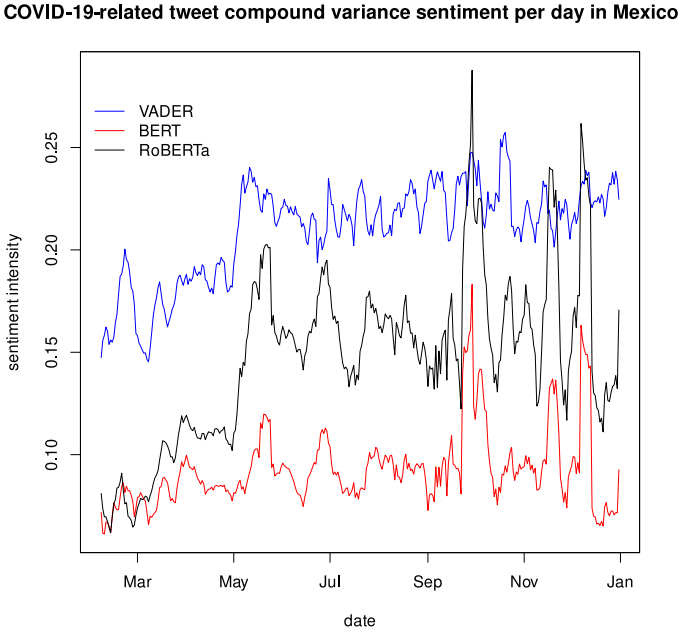

Figure 8. *Year–Long time series of daily averaged compound sentiment polarity variance of COVID-19-related tweets in Mexico*. Each line represents a time series based on the same dataset collected [6] and restricted to Mexico, from 2 February to 31 December 2020. Data were smoothed via a 7-day rolling mean.

Another exciting result lies in the Pearson's correlation test, which shows a strong correlation between the models, as is shown in Table 5. This suggests temporal stability in

the language models and that their difference lies in the magnitude of the sentiment polarity. Of course, this is the expected behavior as we are analyzing the same Twitter corpus. There is no reason why a particular language model would cause temporal warping in the time series. As summarized in Figures 7 and 8, there is an offset in magnitude between the different language models. The same is true of the variance. Regarding the ARIMA models, we have a good fit for all language models with a $p\text{-}value < 2.2 \times 10^{-16}$, and a robust weekly seasonality is present in the data. The charts for these models can be found in the Appendix A. As expected, the trend is almost 0 with $m = 0.00001110643$ for VADER, $m = 0.0004020335$ for BERTweet, and $m = 0.0006753483$ for RoBERTa.

These results make the different language models analyzed for this particular sizeable twitter-based task similar in terms of robustness and stability but not so in terms of sentiment polarity intensity. However, this last metric can be compensated for by normalizing all three results and adjusting the results gathered by VADER, though better results are obtained by utilizing either the positive or negative polarity. Given these minor differences but the similar trend, seasonality, and event reaction, we can draw similar conclusions from this data analysis project, making the language models interchangeably. The only real difference lies within the magnitude of the variance of the sentiment polarities, which should be considered when deciding on any given language model.

5. Discussion

We performed a sentiment analysis on a large COVID-19-only Twitter-based corpus [6], restricted to Mexico, and ranging from February to December 2020, collecting a total of n = 760,064,879 unique tweets. To enable this study, we implemented a flexible software architecture based on cloud-native and serverless technologies to adjust the scale and handle large datasets and streaming data while allowing one to swap out different language models promptly and effortlessly. This solution utilized micro triggers to produce and process a data stream with the modularity of a single tweet, thus allowing one to change data sources and provide the sentiment polarity in a near-real-time fashion. The same was valid for the preprocessing and a portion of the analysis made. We followed the same time series analysis to compare the impact of choosing VADER, BERT, or RoBERTa, as the measurement instrument for determining the sentiment polarity.

To safeguard the users' privacy and comply with international and Twitter's privacy policy, we stripped down all data from personally identifiable information and any metadata, regardless of whether it could be used against a user. To this end, the final schema only had two points, the raw tweet text and the timestamp, adding the needed sentiment polarity calculations for the multiple language models utilized. Doing so did impose some limitations on the data analysis we were able to perform. For example, [16] ranked users and followed the trends and topics for the most popular users, regardless of whether it was COVID-19-related or not. This was not possible for us, as we did not keep users' IDs, nor did we consume tweets that had nothing to do with the COVID-19 pandemic. Another self-imposed limitation of the study was to consider data only from the geographical location of Mexico and only for the year 2020. This was chosen to better align with a previous study [2] and due to a limitation of resources since we preferred to analyze a large data corpus of n = 760,064,879 and have multilanguage data there. Another limitation that can be addressed in the future was the inclusion of other families of language models. We only selected the top state-of-the-art architectures in terms of performance, but it would be interesting to see how other language models, particularly ones with different implementations, perform when executing this extrinsic task. Finally, other authors have performed an additional analysis using as a base the sentiment analysis, such as correlating it to following social distancing during the COVID-19 pandemic [27]; however, this study was limited to the sentiment polarity time series analysis and the language model comparisons.

We provided a time-series analysis of n = 2,142,800 comparing the results of different, popular, state-of-the-art language models reusing a methodology that relied solely on the VADER implementation. The results showed a better stability of modern architectures,

especially BERT, which its training dataset could explain. While BERTweet uses a general, curated corpus, RoBERTa utilizes a Twitter-based corpus for fine-tuning, allowing BERTweet to handle better text outside the known distribution. On the contrary, the magnitude of the sentiment polarities varied slightly. Modern architectures showed a more vital polarity, but when adjusting for VADER's distribution estimation for normalizing results, VADER and RoBERTa produced similar results, with smaller peaks observed in the latter. However, all models showed similar trends and reacted similarly to real-world events, making all three good options for large-scale sentiment analysis systems. RoBERTa is built with 355M parameters, making both forward and backward propagation an expensive operation. In contrast, VADER being a rule-based language model provides a faster, less-expensive solution.

6. Conclusions

We performed a sentiment analysis on a large COVID-19-only Twitter-based corpus, which we referred to as the COVID-19 Twitter chatter dataset [6], but consumed it with a geolocation restricted to Mexico, and ranging from February to December 2020, collecting a total of n = 760,064,879 unique tweets. To enable this study, we implemented a flexible software architecture based on cloud-native and serverless technologies to adjust the scale and handle large datasets. This solution utilized micro triggers to produce and process a data stream with the modularity of a single tweet, thus allowing one to change data sources and provide the sentiment polarity in a near-real-time fashion. The same was valid for the preprocessing and a portion of the analysis made.

We produced a public dataset of a multilanguage sentiment analysis of COVID-19-related tweets in Mexico, utilizing multiple language models as measurement instruments and providing insights on the differences in results in each of them. Although there were differences in the results provided, they were comparable. We remarked that they could be safely used in this study, highlighting the differences in sentiment polarities, stability, and evaluation cost for each. This can be used as a guideline for the technology to be used as well as provide clear insights into the behavior of these models in the real world, keeping in mind that, for the most part, the performance of a language model in a specific extrinsic task is still unclear when only taking into account the intrinsic measurements of each model.

Author Contributions: Conceptualization, E.L.-S. and L.I.B.-S.; methodology, E.L.-S. and L.I.B.-S.; writing—original draft preparation, E.L.-S.; writing—review and editing, L.I.B.-S.; supervision, M.Z. and L.E.F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by Sciences Reseach Council (CONACyT) through the scholarship (grant for CVU 1006856).

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to the open data stated in Twitter's data privacy policy (https://twitter.com/en/privacy).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available upon request to the corresponding author.

Acknowledgments: This research was supported by the Sciences Research Council (CONACyT).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation function
API	Application programming interface
ARIMA	Autoregressive integrated moving average
BERT	Bidirectional Representation for Transformers
CONACyT	Sciences Research Council
COVID-19	Coronavirus disease 2019
CPU	Central processing unit
GCP	Google Cloud Platform
GLUE	General Language Understanding Evaluation
GPU	Graphics processing unit
MLops	Machine learning operations
NLP	Natural language processing
RoBERTa	Robustly optimized BERT pretraining approach
TimeLM	Time language model
VADER	Valence Aware Dictionary for Sentiment Reasoning

Appendix A

We present additional graphs and tables showcasing noncritical information on the sentiment polarity time series for the multiple languages analyzed in this research work.

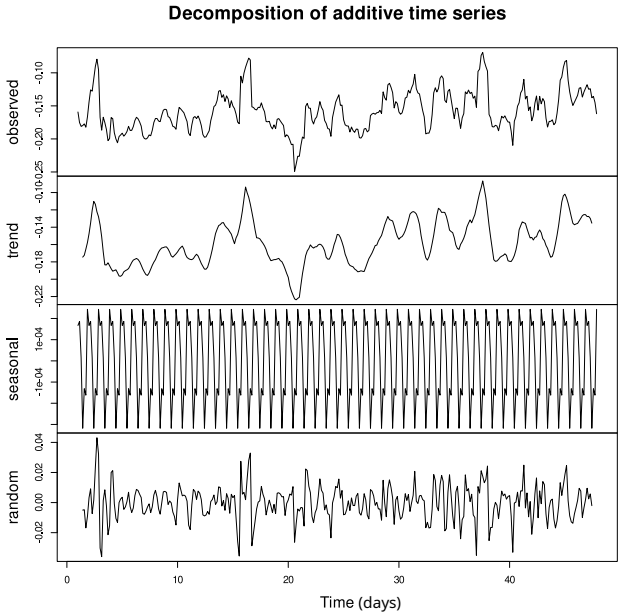

Figure A1. *Year–Long time series of daily averaged compounded sentiment polarity of COVID-19-related tweets, utilizing VADER in Mexico*. Shown is the decomposed sentiment analysis: the averaged time series, trends, weekly seasonality, and random data. The time axis is shown in days, starting from 2 February up to 31 December 2020.

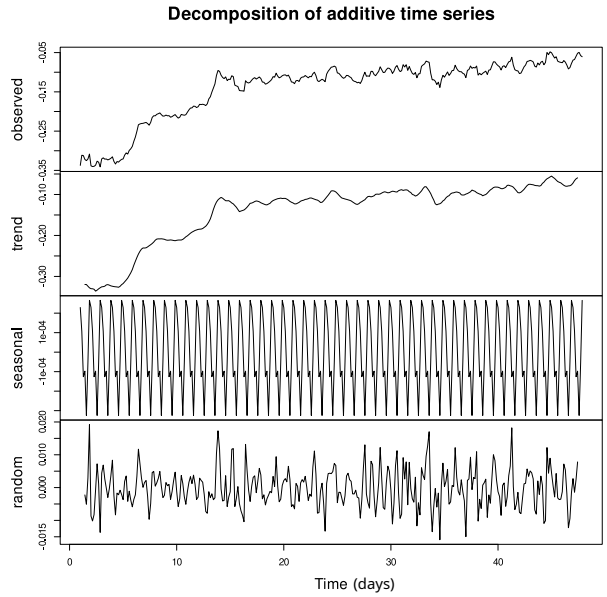

Figure A2. *Year–Long time series of daily averaged compounded sentiment polarity of COVID-19-related tweets, utilizing RoBERTa in Mexico*. Shown is the decomposed sentiment analysis: the averaged time series, trends, weekly seasonality, and random data. The time axis is shown in days, starting from 2 February up to 31 December 2020.

Figures 6, A1 and A2 showcase a summary of the ARIMA model for each of the language models analyzed in this study, ranging from 2 February to 21 December 2020, with a daily unit of time in the x axis. The time series were smoothed using a 7-day rolling average, and we can observe similar trends between BERTweet and RoBERTa and similar seasonality in all of the models. The ARIMA models present a $p\text{-}value < 2.2 \times 10^{-16}$ for all of them, and the trend is almost 0 with $m = 0.00001110643$ for VADER, $m = 0.0004020335$ for BERTweet, and $m = 0.0006753483$ for RoBERTa.

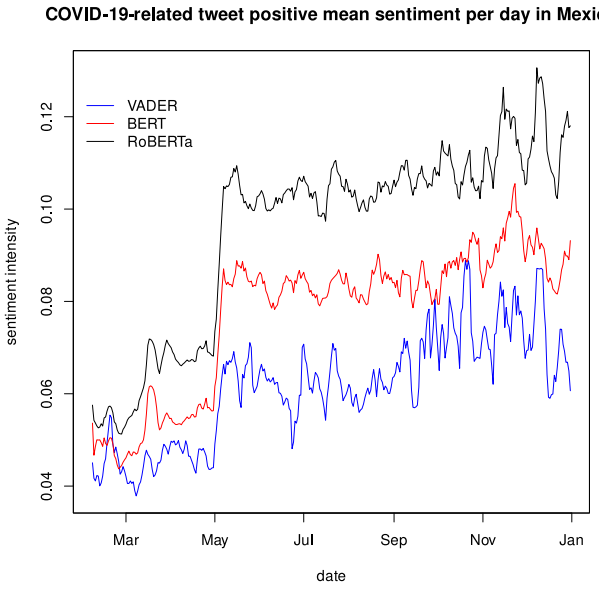

Figure A3. *Year–Long time series of daily averages of positive sentiment polarity of COVID-19-related tweets in Mexico*. Each line represents a time series based on the same dataset collected [6] and restricted to Mexico, from 2 February to 31 December 2020. Data were smoothed via a 7-day rolling mean.

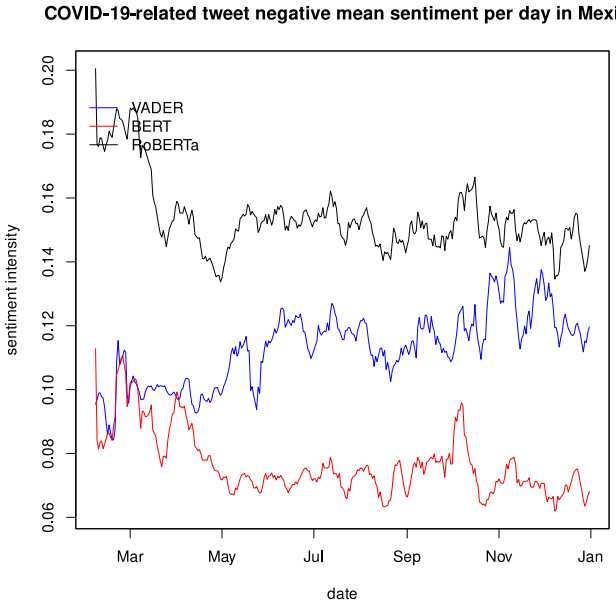

Figure A4. *Year–Long time series of daily averaged negative sentiment polarity of COVID-19-related tweets in Mexico.* Each line represents a time series based on the same dataset collected [6] and restricted to Mexico, from 2 February to 31 December 2020. Data were smoothed via a 7-day rolling mean.

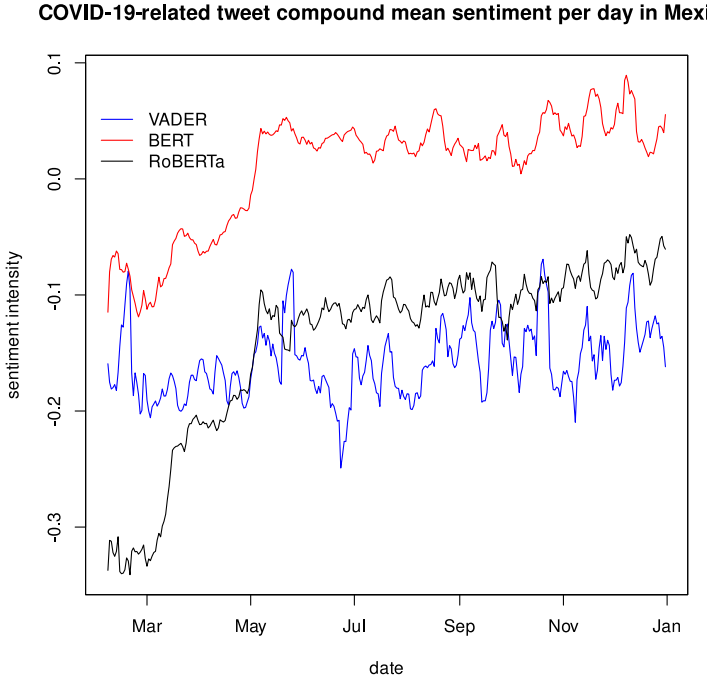

Figure A5. *Year–Long time series of daily averaged compound sentiment polarity of COVID-19-related tweets in Mexico.* Each line represents a time series based on the same dataset collected [6] and restricted to Mexico, from 2 February to 31 December 2020. Data were smoothed via a 7-day rolling mean.

Figures A3–A5 show the daily average of positive, negative, and compound sentiment polarity from 2 February to 31 December 2020. Please note that the *compound* component was calculated using the same α as VADER used for that data entry and is defined by Equation (2); this was to make a similar comparison between the data. All the time series were smoothed over a 7-day rolling average, ranging from $(0, 1)$ for the positive and negative polarities, and $(-1, 1)$ for the compound polarity. Note that in the positive

polarity, VADER is the least intense, followed by BERT and, lastly, by RoBERTa, while in the negative polarity, BERT is the least intense, followed by VADER, and then by RoBERTa.

Table A1 summarizes the partial autocorrelation function indexes. We observed strong indications of weekly autocorrelations, which helped us quickly find the correct coefficients for fitting an ARIMA model and use it to decompose the time series.

Table A1. Partial ACF (auto correlation function) of the sentiment polarity time series.

Lag	Partial ACF		
	VADER	BERTweet	RoBERTa
1	0.888493947	0.981322253	0.9830592776
2	−0.188964510	0.058733847	0.0372013134
3	−0.119674873	−0.009975352	−0.0251595535
4	−0.030455558	−0.010288171	−0.0380042091
5	−0.059030857	−0.071735183	−0.0150139920
6	−0.009066172	−0.021977650	0.0105725987
7	−0.074336784	−0.006658723	0.0308749961
8	0.483162228	0.070486755	0.0321131066
9	−0.174833350	0.018110652	0.0085143348
10	−0.106022420	−0.026950394	−0.0216328943
11	−0.042640744	−0.016666002	−0.0391064396
12	0.036015376	0.054023430	0.0005282503
13	−0.001715747	0.057808036	−0.0174470950
14	−0.022592658	0.041424359	−0.0488539568
15	0.262433976	−0.018884831	0.0616588475
16	−0.113870361	−0.025551679	−0.0041771775
17	0.014280624	−0.030573729	−0.0247434641
18	−0.095572696	−0.028879640	−0.0417809239
19	0.040884337	−0.060802245	−0.0123302508
20	−0.050324985	−0.023285051	−0.0306742896
21	0.007294918	−0.019094493	−0.0023443146
22	0.183424253	0.060710941	−0.0083751616
23	−0.077040062	−0.012948749	−0.0258095563
24	−0.088658363	−0.045977837	−0.0529454175
25	−0.023658966	−0.044841702	0.0490646129

References

1. Huerta, D.T.; Hawkins, J.; Brownstein, J.; Hswen, Y. Exploring discussions of health and risk and public sentiment in MA during COVID-19 pandemic mandate implementation: A Twitter analysis. *SSM-Popul. Health* **2021**, *15*, 100851. [CrossRef] [PubMed]
2. León-Sandoval, E.; Zareei, M.; Barbosa-Santillán, L.I.; Falcón Morales, L.E.; Pareja Lora, A.; Ochoa Ruiz, G. Monitoring the Emotional Response to the COVID-19 Pandemic Using Sentiment Analysis: A Case Study in Mexico. *Comput. Intell. Neurosci.* **2022**, *2022*, 4914665. [CrossRef] [PubMed]
3. El Alaoui, I.; Gahi, Y.; Messoussi, R. Full Consideration of Big Data Characteristics in Sentiment Analysis Context. In Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 12–15 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 126–130.
4. Laney, D. 3D data management: Controlling data volume, velocity and variety. *META Group Res. Note* **2001**, *6*, 1. Available online: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed on 23 June 2022).
5. Ylijoki, O.; Porras, J. Perspectives to Definition of Big Data: A Mapping Study and Discussion. *J. Innov. Manag.* **2016**, *4*, 69–91. [CrossRef]
6. Banda, J.M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, K.; Tutubalina, E.; Chowell, G. A large-scale COVID-19 Twitter chatter dataset for open scientific research—An international collaboration [DataSet]. *Epidemiologia* **2021**, *2*, 315–324. [CrossRef]
7. Cenni, D.; Nesi, P.; Pantaleo, G.; Zaza, I. Twitter vigilance: A multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis. In *Proceedings of the 2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017, San Francisco, CA, USA, 4–8 August 2017*; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8. [CrossRef]

8. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8.
9. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; Association for Computational Linguistics: Vancouver, BC, Canada, 2020. Available online: https://aclanthology.org/2020.emnlp-demos.2/ (accessed on 23 June 2022).
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; MAG ID: 2896457183.
11. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
12. Loureiro, D.; Barbieri, F.; Neves, L.; Anke, L.E.; Camacho-Collados, J. TimeLMs: Diachronic Language Models from Twitter. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Dublin, Ireland, 22–27 May 2022. Available online: https://aclanthology.org/2022.acl-demo.25/ (accessed on 23 June 2022).
13. Barbieri, F.; Camacho-Collados, J.; Neves, L.; Espinosa-Anke, L. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics*; Association for Computational Linguistics: Vancouver, Canada, 2020. Available online: https://aclanthology.org/2020.findings-emnlp.148/ (accessed on 23 June 2022).
14. Wan Min, W.N.S.; Zulkarnain, N.Z. Comparative Evaluation of Lexicons in Performing Sentiment Analysis. *J. Adv. Comput. Technol. Appl.* **2020**, *2*, 14–20.
15. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, August 2017; Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 502–518. [CrossRef]
16. Adikari, A.; Nawaratne, R.; De Silva, D.; Ranasinghe, S.; Alahakoon, O.; Alahakoon, D. Emotions of COVID-19: Content Analysis of Self-Reported Information Using Artificial Intelligence. *J. Med. Internet Res.* **2021**, *23*, e27341. [CrossRef]
17. Lwin, M.O.; Lu, J.; Sheldenkar, A.; Schulz, P.J.; Shin, W.; Gupta, R.; Yang, Y. Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends. *JMIR Public Health Surveill.* **2020**, *6*, e19447. [CrossRef]
18. Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hamdi, M.; Shah, Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *J. Med. Internet Res.* **2020**, *22*, e19016. [CrossRef]
19. Boon-Itt, S.; Skunkan, Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health Surveill.* **2020**, *6*, e21978. [CrossRef]
20. Xue, J.; Chen, J.; Hu, R.; Chen, C.; Zheng, C.; Su, Y.; Zhu, T. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *J. Med. Internet Res.* **2020**, *22*, e20550. [CrossRef]
21. Valdez, D.; ten Thij, M.; Bathina, K.; Rutter, L.A.; Bollen, J. Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *J. Med. Internet Res.* **2020**, *22*, e21418. [CrossRef]
22. Crocamo, C.; Viviani, M.; Famiglini, L.; Bartoli, F.; Pasi, G.; Carrà, G. Surveilling COVID-19 Emotional Contagion on Twitter by Sentiment Analysis. In *European Psychiatry*; Cambridge University Press: Cambridge, UK, 2021; Volume 64, p. e17. [CrossRef]
23. Chandra, R.; Krishna, A. COVID-19 sentiment analysis via deep learning during the rise of novel cases. *PLoS ONE* **2021**, *16*, e0255615. [CrossRef]
24. Alam, K.N.; Khan, M.S.; Dhruba, A.R.; Khan, M.M.; Al-Amri, J.F.; Masud, M.; Rawashdeh, M. Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data. *Comput. Math. Methods Med.* **2021**, *2021*, 4321131. [CrossRef]
25. Garcia, K.; Berton, L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Appl. Soft Comput.* **2021**, *101*, 107057. [CrossRef]
26. Singh, M.; Dhillon, H.; Ichhpujani, P.; Iyengar, S.; Kaur, R. Twitter sentiment analysis for COVID-19 associated mucormycosis. *Indian J. Ophthalmol.* **2022**, *70*, 1773. [CrossRef]
27. Porcher, S.; Renault, T. Social distancing beliefs and human mobility: Evidence from Twitter. *Plos ONE* **2021**, *16*, e0246949. [CrossRef]
28. Shofiya, C.; Abidi, S. Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5993. [CrossRef]
29. Jaidka, K.; Giorgi, S.; Schwartz, H.A.; Kern, M.L.; Ungar, L.H.; Eichstaedt, J.C. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 10165–10171. [CrossRef]
30. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv* **2018**, arXiv:1804.07461. [CrossRef]
31. Kmetty, Z.; Koltai, J.; Bokányi, E.; Bozsonyi, K. Seasonality pattern of suicides in the US—A comparative analysis of a Twitter based bad-mood index and committed suicides. *Intersect. East Eur. J. Soc. Politics* **2017**, *3*, 56–75. [CrossRef]
32. Müller, M. Dynamic time warping. *Inf. Retr. Music. Motion* **2007**, 69–84.