Impact of Covid-19 on passengers and airlines from passenger measurements: Managing customer satisfaction while putting the US Air Transportation System to sleep

Philippe Monmousseau,⁎ Aude Marzuoli, Eric Feron, Daniel Delahaye

ENAC, Université de Toulouse, France
Division of Electrical, Computer and Mathematical Science and Engineering, King Abdullah University of Science and Technology

ABSTRACT

The COVID-19 pandemic has had a significant impact on the air transportation system worldwide. This paper aims at analyzing the effect of the travel restriction measures implemented during the COVID-19 pandemic from a passenger perspective on the US air transportation system. Four metrics based on data generated by passengers and airlines on social media are proposed to measure how the travel restriction measures impacted the relation between passengers and airlines in close to real-time. The proposed metrics indicate that each airline has reacted differently to the COVID-19 travel restriction measures from a passenger perspective, therefore they can be used by airlines and passengers to improve their decision making process. This report comes ahead of official data related to the same sequence of events, thereby showing the value of passenger-borne data in an industry where corporate priorities, institutional prudence, and passenger satisfaction come close together.

Keywords: Air transportation system, Passenger-generated data, Passenger-centric metrics, COVID-19

1. Motivation

1.1. The COVID-19 pandemic and the resulting travel restrictions from a US perspective

In response to the pandemic situation resulting from the outbreak of the corona disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), travel restrictions measures were implemented by various countries, impacting both domestic travel and international travel (New York Times, 2020). Italy was the first country to enforce a national lockdown (WorldAtlas, 2020) on March 9th 2020, after introducing on February 21st 2020 an initial measure confining only the northern region of Lodi. Two days after Italy's lockdown announcement, on March 11th 2020, the United States banned non-US travelers who had been to China, Iran and 26 member states of the European Union (EU) to enter the US, and later extended the ban to non-US travelers who had visited the United Kingdom and Ireland on March 16th 2020 (New York Times, 2020). The EU officially closed the external borders of 26 of its member states to nearly all non-EU residents on March 17th 2020 (New York Times, 2020). On March 19th 2020, the US Department of State issued a Level 4 Global Health Travel Advisory, which cautions all US citizens against international travel, still in place as of May 6th 2020 (US Department of State, 2020).

This dramatic sequence of events forms the thread against which the air transportation system has had to progressively put itself to a semi-comatose state to address fast-growing sanitary and economic concerns. For these reasons, the following dates are indicated with dotted lines in every graph throughout this paper in order to better visualize the timeline of each figure.

1. The Lodi region lockdown in Italy: February 21st, 2020
2. Italy's lockdown: March 9th, 2020
3. US ban of non-US travelers from the EU, China and Iran: March 11th, 2020
4. EU external border closure: March 17th, 2020
5. US Level 4 Global Health Travel Advisory: March 19th, 2020.

Fig. 1 presents the number of passengers arriving at US immigration across all airports of entry using the “Airport Wait Times” data from the Customs and Border Protection (CBP) website (United States Customs and Border Protection, 2020). This plot illustrates clearly the effect of these travel restriction measures on the international traffic coming to the US. For a more detailed presentation of the available CBP dataset, the authors recommend reading Monmousseau et al. (2019a), which also presents an analysis of the wait times at US airport immigration services from January 2013 to January 2019.

⁎ Corresponding author.
E-mail address: philippe.monmousseau@enac.fr (P. Monmousseau).

http://dx.doi.org/10.1016/j.trip.2020.100179
2590-1982/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
The air transportation system is an essential system to understand and to study under pandemic situations from various perspectives, e.g. the propagation of diseases inside airplanes (Namiala et al., 2017), the propagation of epidemics via airplanes (Chinazzi et al., 2020), or the effect of the travel restrictions on airline employment (Sobieralski, 2020). This paper focuses on the effect of the pandemic on the attitude of passengers towards airlines. In 2019, considering eight major US airlines and thirty-four major US airports, Twitter users wrote a median of 13,255 tweets mentioning an airport and a median of 295,904 tweets mentioning an airline, indicating that users interact more with airlines than with airports.

1.2. The limitations of traditional approaches to assess the impact of COVID-19 on the air transportation system

The travel restrictions, and the other measures taken by a majority of countries worldwide, are having an unprecedented impact on the air transportation system. Until official flight data are released in the United States regarding international and domestic air transportation there are no means of measuring this impact on the US air transportation system, except by relying on non-traditional data sources.

Traditionally, the metrics used to measure the state of the US air transportation system are focused on flight performances, such as the amount of delay per flight, the number of delayed flights, the number of cancelled flights and the number of carried passengers. The data considered for these metrics are gathered by the US Department of Transportation Bureau of Transportation Statistics (BTS) (Bureau of Transportation Statistics, 2018). The data are first processed by airlines and airports and then provided to the BTS, which then publishes the data as a monthly report. The BTS reports pertaining to on-time flight data are usually published with a latency of two months. This latency is not well adapted for monitoring and analyzing the effects of situations such as the COVID-19 pandemic on the US air transportation system.

Fig. 2 presents the number of international and domestic flights from March 1st 2020 to April 22nd 2020 using available data from CBP and from BTS as of June 24th 2020. From these data, the number of daily domestic flights drops by half in the second half of March 2020 but no conclusion can be drawn for the month of April 2020. While not technically asleep, many flights kept being flown by airlines because they feared to lose their slots (Truax, 2020; Business Insider, 2020) or because they had to keep flying routes in order to receive financial aid (Congress of the United States of America, 2020), a situation close to a sleep condition called “nightmare”.

Additionally, various studies have shown that passengers were disproportionately impacted by flight capacity reduction (Bratu et al., 2005; 2006; Wang et al., 2006; Wang, 2007), highlighting the differences between measuring flight delays and flight cancellations and measuring the actual passenger delay. For example, based on data from a major US airline, they show that disrupted passengers, whose journey was interrupted by a capacity reduction, are only 3% of the total passengers, but suffer 39% of the total passenger delay. The necessity of adding a passenger-centric approach when evaluating the air transportation system was later put forward by NextGen (Gawdiak and Diana, 2011) in the US and by ACARE FlightPath 2050 in Europe (Darecki et al., 2011). A first attempt at implementing passenger-oriented metrics was performed by Cook et al. (Cook et al., 2012). Integrating passenger objectives in airport decision making processes was introduced within the concept of Multimodal, Efficient Transportation in Airports and Collaborative Decision Making (META-CDM) (Laplace et al., 2014; Kim et al., 2013; Dray et al., 2015). Though these works give an important place to passengers, they still heavily rely on flight-centric data and have thus the same latency limitation.

Several years later, the advocated shift from flight-centric metrics to passenger-centric metrics still has to be actually implemented by the governing agencies. In a report published in 2016, EUROCONTROL and the FAA presented metrics regarding punctuality that combines airline and passenger views into a single view (EUROCONTROL and Federal Aviation Administration Air Traffic Organization System Operations Services, 2016).

Already in 1992, Lemer (1992) advocated for the need of unified airport performance measures that would balance the expectations of passengers, airlines and airports along with the expectations of other actors (such as restaurants or governments). Understanding the passenger experience, or at least the passenger perception of airport and airline quality has since been the focus of many studies. Tsaur et al. (2002) first proposed to introduce surveys based on fuzzy set theory in order to analyze airline service quality. Hunter (2006) performed a thorough survey of airline perception related studies from 1995 to 2006, pointing out the decrease in customer service throughout the airline industries. For more informations on the various survey-based methods used, de Oña and de Oña (2015) conducted a survey of survey based analysis of public transportation system. They concluded that even though researchers keep trying to improve the complexity of the models to better model passenger satisfaction of a public transportation system, managers and practitioners use simpler models in order to reach their goal of improving passenger perceived service quality for an increase of income.

Passenger surveys conducted at airports for airports or airlines, while very detailed, remain limited to very small samples of passengers and short time periods, and may not be representative. For example, among some of the founding survey studies, Tsaur et al. (2002) have a sample size of 211 passengers and Pakdil and Aydin (2007) have a sample size of 385 passengers. They are also expensive and time consuming to implement, making their use for measuring the effects of major perturbations, such as the COVID-19 pandemic, on the air transportation system cumbersome and difficult to update.

1.3. Passengers as sensors of the air transportation system

Using passenger-generated data in order to analyze the efficiency of the air transportation system was made easier thanks to the ubiquity of smartphones. Data from WiFi hotspots and Bluetooth beacons, along with historical data, are used to analyze passenger behavior at airports (Nikouei et al., 2015; Huang et al., 2019) and at transit stations (Van den Heuvel et al., 2011). If available, data generated by passengers smartphone and collected by phone carriers can be processed to analyze the door-to-door service behavior of passengers (Marzuoli et al., 2019, 2018; García-Albertos et al., 2017), both under nominal and degraded conditions. However data gathered directly from smartphones are proprietary data and are not often publicly available for research.

This paper proposes an alternative approach to analyzing the air transportation system by focusing on airline performances with respect to their passengers using data generated by airlines and by passengers. The importance of airlines of improving the waiting environment at airports in order to improve passenger satisfaction is already highlighted in Pruyn and Smidts (1998) and is generalized for riders at transit stations in Watkins et al. (2011). In the specific case of US air transportation, Twitter is an important medium for direct communication between passengers and airlines.
are determined based on a set of manually de-
to complaints than to compliments, where complaints and compliments

customers with greater popularity, and have a tendency to respond more
towards which airline services are associated with positive or negative sentiments.

tories (personal, positive, negative, promotion, question or news). Tweets

constructed from passenger-generated data in order to offer a

tness is investigated in Monmousseau et al. (2019c,b) whose purpose is to es-
mored keywords. These studies

and “good” and “fail”-ing an equal number of tweets expressing a

expectation is embedded in the tweet. To remove any potential bias of emojis on

depending on the works of Read (2005); Go et al. (2009). Emoji filters are

used to extract 49,030 tweets written in 2017 by airlines and their cus-

m more than 75% of the tweets within the training dataset.

ctor, and are therefore inherently biased, towards both extreme
dissatisfaction and extreme satisfaction. However, Sampson (1996) sug-
suggests that continuous quality monitoring can benefit from the extreme re-
sponses contained in unsolicited feedback. For example, unsolicited

feedback within social media activity is used by Abrahams et al. (2013) to

detect information about defective components in the automotive indus-

try. In Europe, KLM promised a 30-minute customer-response time in the

aftermath of the air transportation major disruption initiated by the erup-
tion of an Icelandic volcano in 2010 (Kane, 2014).

The real-time availability of Twitter data is the starting point of many

studies of large scale events, such as natural disasters, and how Twitter
could be used to help emergency responders (Palen et al., 2016; Vieweg

et al., 2010; Kireyev et al., 2009; Terpstra and Stronkman, 2012; Priya

et al., 2020; Srivastava and Sankar, 2020). Regarding applications to the

air transportation field, most works mining Twitter data focus on creating

and improving airline sentiment classification methods (Breen, 2012; 

Wan and Gao, 2015), without proposing any direct use of their results to

improve airline service or passenger satisfaction.

Applications of sentiment analysis for airlines are proposed by Siau

(2014) who use sentiment and topic analysis to extract from around a thou-
sand tweets the information needed to calculate a proxy of the Airline Qual-

ity Rating, a flight-centric metric including a measure of customer

complaints introduced by Bowen et al. (1991). Misopoulos et al. (2014) an-

alyze airline customer service experiences both by manually labelling
tweets related to four airlines written on five different days of 2010 and

containing one of three keywords (“good”, “fail” and “lounge”) into six cat-

ergies (personal, positive, negative, promotion, question or news). Tweets

within the positive and negative categories are then analyzed to determine which

airline services are associated with positive or negative sentiments. 

Gunarathe et al. (2015) show that airlines are more likely to respond to

customers with greater popularity, and have a tendency to respond more to

complaints than to compliments, where complaints and compliments are
determined based on a set of manually defined keywords. These studies show 

that Twitter can be used by airlines in order to gain some insights on

how passengers perceive their service and how they treat their passengers on

Twitter.

Using Twitter to build a real-time estimator of the air transportation sys-
tem is investigated in Monmousseau et al. (2019c,b) whose purpose is to es-
timate flight-centric values per airport before they were released by BTS.

This paper takes another approach and proposes several passenger-centric

metrics constructed from passenger-generated data in order to offer a

passenger-centric perspective of the air transportation system, with a

focus on the relation between airlines and passengers. This paper is not
directly interested in “classical” measures of performance, such as those of

direct interest to airlines (productivity, profitability) or to Air Navigation

Service Providers (on-time performance and other metrics used to evaluate

technical development programs, such as NextGen in the US or SESAR in

Europe). The proposed work introduces measures of satisfaction and feed-

back expressed by the passengers themselves. Such measures are comple-

mentary with and different from the foregoing, although correlations may

exist.

Eight airlines, and their associated Twitter handles, are considered in

the analysis below: American Airlines (@AmericanAir), Delta Air Lines

(@Delta), United Airlines (@united), Alaska Airlines (@AlaskaAir), South-

west Airlines (@SouthwestAir), JetBlue Airways (@JetBlue), Spirit Airlines

(@SpiritAirlines) and Frontier Airlines (@FlyFrontier and @FrontierCare).

The first four are legacy airlines, and the last four are low-cost carriers. All
tweets written from these airlines Twitter accounts were scraped from Feb-

uary 16th 2020 to May 3rd 2020 and are categorized as “customer service 
tweets”. This category contains both public replies to customers and public

tweets for everyone to read. All tweets written over that same period and

mentioning at least one of the airline handles that was not written from the

corresponding airline’s Twitter account were also scraped and catego-

rized as “passenger tweets”.

The rest of this paper is structured as follows: Section 2 describes the

first two metrics based on a Twitter sentiment analysis and how they can be

used in light of the COVID-19 situation. Section 3 then describes two addi-

tional metrics based on selected keywords and how they can be used to

assess the performance of airline communication during the COVID-19 pan-

demic. Section 4 concludes this paper and discusses future research

directions.

2. Impact of the COVID-19 on airline and passenger mood

2.1. Evaluating the mood expressed in tweets

A first step in sentiment analysis is to create a labelled dataset contain-
ing an equal number of tweets expressing a positive sentiment and tweets ex-

pressing a negative sentiment. The training dataset used in this study is

based on the works of Read (2005); Go et al. (2009). Emoji filters are

used to extract 49,030 tweets written in 2017 by airlines and their cus-

omers and automatically assign a positive or negative sentiment label to
each tweet according to Table 1.

A processing pipeline is then applied to each tweet in order to transform 

the text contained within each tweet into a vector of tokens that can be

processed by the sentiment classifiers. A token is either a generic keyword, 
a single word, or a bigram or a trigram. Bigrams (resp. trigrams) are combina-

tions of two (resp. three) consecutive words that are commonly used to-
gether, e.g. “do not like” is a trigram. In order to reduce the sparsity of the

considered vocabulary, generic keywords are used to replace mentions to

other Twitter users (“@someone” becomes “MENTION”) and mentions to

the considered airlines (e.g. “@united” becomes “AIRLINE”). They are also

used to replace date related association of words, e.g. “March 11th

2020” becomes “DATE” and “8 am” becomes “TIME”. Generic keywords are

also used to indicate if the tweet contains a link to a website or if a picture

is embedded in the tweet. To remove any potential bias of emojis on the

learning process, since every tweet in the training dataset contains an

emoji, all emojis are replaced by the keyword “EMOJI”.

Since the text contained in a tweet can be loosely written, with emphasis
given to words with repeated letters such as “loooove”, the number of du-

plicate letters was limited to two in every word: “loooove” becomes

“loooove”. Negative bigrams are also automatically created by merging nega-
tion words (“no”, “not” and “never”) with the word that follows it. Once all 

the tokens were created, the tokens occurring in fewer than twenty tweets

within the training dataset are removed as well as the tokens appearing in

more than 75% of the tweets within the training dataset.

Five classifiers are trained on the training dataset and then tested on the

labelled dataset provided by Kaggle (2018): an AdaBoost classifier (Freund

and Schapire, 1995), a gradient boosting classifier (Friedman, 2001), a
random forest classifier (Breiman, 2001), a naive Bayesian classifier (Chan et al., 1982) and a logistic regressor (Yu et al., 2011) using the scikit-learn python library (Pedregosa et al., 2011).

Each classifier gives a score of 1 if it considers that the tweet expresses a positive sentiment and a score of 0 if it expresses a negative sentiment. In effect, each classifier calculates a predicted probability for a tweet of being positive, and then rounds that predicted probability to the closest integer (0 or 1). The classifiers are transformed into regressors by considering the probability for a tweet of being classified as positive. The output of all trained regressors is then averaged into one single score ranging from 0 to 1, with a score of 0 indicating a negative mood and a score of 1 indicating a positive mood.

2.2. Daily mood evolution

Once the sentiment expressed within each tweet is averaged on a daily level, the effect of the travel restriction measures on the expressed passenger mood can be compared with their effects on the expressed airline mood. Legacy airlines are usually considered as offering a higher quality service to customers than low-cost carriers, with an average of close to 296 tweets written a day by the customer service of the four considered US legacy airlines versus an average of 112 tweets written a day by the customer service a day for the four considered low-cost carriers. The evolution of the mood expressed by passengers and airline customer services is presented in the following subsections, first for the legacy airlines and then for the low-cost carriers.

2.2.1. Case of legacy airlines

Fig. 3 shows the evolution of the mood expressed by the four legacy airlines considered and by their passengers from February 16th 2020 to May 3rd 2020.

From Fig. 3(a), a drop in the mood expressed by passengers can be observed starting right after the Lodi lockdown with a steep decrease right after the US travel ban for the three major airlines (Delta Air Lines, United Airlines and American Airlines). The sentiment extracted from the tweets from Delta’s passengers has the steepest descent but also the sharpest recovery. The case of Alaska Airlines exhibits special characteristics: a #AlaskaHappyHour campaign, giving Twitter users the opportunity of winning a free flight to Alaska, was taking place early March 2020. This campaign could explain why the expressed mood in passenger tweets increased between March 1st 2020 and March 5th 2020 and could have compensated a potential decrease in the passenger expressed mood linked to the travel ban announcement.

Regarding the mood expressed in tweets written by the airline customer services, shown in Fig. 3(b), it only decreases for Delta Air Lines and United Airlines starting at the announcement of Italy’s lockdown. An opposite reaction is seen with the mood expressed by American Airlines customer service, which increases over that same period. Comparing Fig. 3(a) and Fig. 3(b) shows that Delta Air Lines and Alaska Airlines have the highest expressed mood on average within their passenger tweets over the considered period, but the lowest expressed mood within their customer service tweets of the four legacy airlines. An explanation of the better mood expressed by their passengers could be that these airlines expressed a mood closer to their passengers’ actual mood. A gap between the mood extracted from passenger tweets and the mood extracted from airline customer service tweets is visible from one figure to another, with airline customer service tweets expressing a mood about 0.2 points higher than passenger tweets.

2.2.2. Case of low-cost carriers

Similar conclusions can be drawn when analyzing the mood associated to tweets from passengers and customer services of low-cost carriers. Fig. 4 shows the evolution of the expressed mood from February 16th 2020 and May 3rd 2020 in the passenger and customer service tweets of the three low-cost carriers considered.

Fig. 4(a) indicates that the mood expressed by Spirit Airlines passengers and by Frontier Airlines passengers is significantly lower on average than the mood expressed by passengers of JetBlue Airways and Southwest Airlines over the months of February and March 2020. There is a spike in the mood extracted from tweets written by JetBlue passengers around March 26th 2020. This date is also the day when the governor of New York thanked JetBlue for offering free flights to health care workers in order to help the state handle the spread of COVID-19. It also corresponds to the period when an update of their mobile application contained the message “Now, go wash your hands”, prompting an amused reaction of their passengers. The drop in the mood expressed in the tweets written by legacy airline passengers after Italy’s lockdown is less visible in the tweets written by passengers of low-cost carriers, with the exception of the mood expressed by passengers of Southwest Airlines.

Looking at the mood expressed by low-cost carrier customer services presented in Fig. 4(a), the mood expressed by the customer service of Frontier Airlines displays a highly varying behavior, oscillating between 0.23 and 0.83 with discontinuities since on certain days no tweets were written by their customer service. For the other three low-cost carriers, the gap between the mood extracted from the tweets written by Southwest Airlines customer service and the mood extracted from the tweets written by the customer services of the other two carriers reduces significantly the day after Italy’s lockdown. Similarly as for legacy airlines, a gap of about 0.2 points is visible between the mood expressed within passenger tweets and airline customer service tweets by comparing Fig. 4(a) and (b).

2.3. Passenger-centric metrics

Based on the observations presented in Section 2.2, two passenger-centric metrics are proposed to measure the relation between airline customer services and their passengers. The first proposed metric aims at measuring the evolution of the airline mood relative to the mood of their passengers. Diverging mood evolutions are given a low score: if the average mood expressed by passengers is decreasing, the average mood expressed in the tweets written by the airline customer service should not be increasing.

**Proposed passenger-centric metric 1.** The airline empathy score is defined as the Pearson correlation between the evolution of the average mood expressed by passengers in their tweets and the evolution of the average mood expressed by the airline customer service in their tweets.

The empathy score $\Xi$ is calculated using the following formula:

$$\Xi = \frac{\sum (p_i - \bar{p})(c_i - \bar{c})}{\sqrt{\sum (p_i - \bar{p})^2 \sum (c_i - \bar{c})^2}}$$

where the set $(p_i)$ (resp. $(c_i)$) is the ordered set of the daily expressed mood in passenger tweets (resp. in airline customer service tweets), and $\bar{p}$ (resp. $\bar{c}$) is the average daily expressed mood over the considered period in passenger tweets (resp. in airline customer service tweets).

The empathy score $\Xi$ goes from $-1$ to $1$, with a score of $1$ meaning that the airline customer service expressed mood is in agreement with the mood expressed by their passengers. On the opposite, a score of $-1$ indicates that the mood expressed by the airline customer service is in complete opposition of phase with the mood expressed by their passengers. Such a score would indicate that the mood expressed by the airline customer service increases when the mood expressed in

**Table 1**

| Emoji sentiment association. |
|-----------------------------|
| Category | Emojis |
| Positive | “😊”, “😊😊”, “😊😊😊”, “😊😊😊😊”, “😊😊😊😊😊”, “😊😊😊” |
| Negative | “😊😊😊😊😊”, “😊😊😊😊”, “😊😊😊” |

1 https://twitter.com/NYGovCuomo/status/1242941065535608835.
passenger tweets decreases, and vice-versa. A score of 0 indicates that the mood expressed by the airline customer service and the mood expressed by their passengers are uncorrelated.

The second proposed metric aims at measuring the gap observed between the mood expressed by passengers in their tweets and the mood expressed in the tweets written by airline customer services.

**Proposed passenger-centric metric 2.** The airline sentiment gap is the average difference between the mood expressed by passengers and the mood expressed by airlines.

The airline sentiment gap $\Delta$ is calculated using the following formula:

$$\Delta = \frac{1}{N} \sum_{i} (p_i - c_i)$$

where $N$ is the number of days considered and the set $\{p_i\}$ (resp. $\{c_i\}$) is the ordered set of the daily expressed mood in passenger tweets (resp. in airline customer service tweets), as for the airline empathy score $\Xi$ presented in Eq. (1).

The airline sentiment gap $\Delta$ goes from $-1$ to 1 with a gap of 0 indicating that airline customer services and passengers express the same average mood in their tweets. A gap of 1 indicates a mood expressed by an airline customer service equal to 1 (i.e. the highest possible mood) and a mood expressed by the airline passengers equal to 0 (i.e. the lowest possible mood) on every day of the considered period. A gap of $-1$ indicates the opposite scenario.

Table 2 shows the ranks and scores of the seven airlines associated with each of the two passenger-centric metrics proposed in this section. Both the empathy score $\Xi$ and the sentiment gap $\Delta$ were calculated over the period from March 1st 2020 to March 31st 2020.

### 3. Keyword-based metrics

#### 3.1. Cancellations

When some exceptional situation occurs, an important increase in the use of specific keywords within the stream of tweets written by the affected users can take place. For example, if many cancellations occur, many passengers will connect to Twitter and write tweets containing the keyword “cancel” to express their concerns directly to the airline they have bought tickets from. In this analysis, any word starting with the keyword “cancel”, such as “cancellation” or “cancelled”, is considered as a keyword “cancel”.

Fig. 5 shows the evolution of the normalized number of tweets written by passengers and containing the keyword “cancel” between February 16th 2020 and May 3rd 2020 for four US legacy airlines and four US low-cost carriers. The normalization is based on the total number of passengers carried by each airline in 2018 and available in the yearly BTS reports Bureau of Transportation Statistics (2020).

Fig. 5(a) indicates that the passengers of the four legacy airlines react as early as Italy’s lockdown announcement with an important increase in the number of tweets containing the keyword “cancel”. A second spike in the number of passenger tweets containing the keyword “cancel” then occurs once the US announces that it bans all travelers from the EU, China and Iran. Fig. 5(a) shows that Delta Air Line passengers were, in proportion, about three times more vocal about cancellations on Twitter than the other legacy airlines at this period. This could be an indication that Delta Air Line had a greater proportion of passengers traveling within or through the EU at that time. The number of tweets from Alaska Airlines passengers containing the keyword “cancel” had an early spike compared to the tweets written by passengers from the other legacy airlines. That early spike could be linked to the fact that most of the early US cases of COVID-19 were
discovered on the US West Coast first, which is where the main hub of Alaska Airlines is located.

Fig. 5(b) shows the evolution of the number of tweets containing the keyword “cancel” written by passengers of the four low-cost carriers. Southwest Airlines passengers were, in proportion, less vocal on Twitter on the matter of cancellation than passengers of the other low-cost carriers, with a slight increase in the number of tweets containing the keyword “cancel” that is almost entirely contained within the period between the announcement of Italy’s lockdown and the start of the US Level 4 Global Health Travel Advisory. JetBlue Airways passengers display a behavior similar to passengers of legacy airlines in this case. Passengers of Spirit Airlines and Frontier Airlines waited until the US travel ban announcement to communicate massively on Twitter their concerns using the word “cancel”. The second spike in the number of tweets containing the keyword “cancel” starting at the announcement of the EU border closure is more important and lasts longer for tweets written by passengers of Frontier Airlines.

Fig. 6 shows the evolution of the number of tweets containing the keyword “cancel” and written by airline customer services between February 16th 2020 and May 3rd 2020 for the same four US legacy airlines and three US low-cost carriers. Please note that the y-axis scale is different in Fig. 6(a) and (b).

Regarding tweets written by legacy airline customer services, the evolution of the number of tweets containing the keyword “cancel” shown in Fig. 6(a) presents similarities for three of the four airlines. There is a significant increase in the number of customer service tweets containing the keyword “cancel” starting the day Italy announced its lockdown and then a slow decrease. For tweets written by American Airlines customer service, the number of tweets containing the keyword “cancel” increases as for the other three airlines, but it does not decrease afterwards but fluctuates at a level more important than during the period before the travel restriction measures were announced.

Regarding low-cost carriers, Fig. 6(b) shows that each carrier use the keyword “cancel” on different occasions. The number of occurrences of the keyword “cancel” within tweets written by Southwest Airlines passengers has two important spikes around each of the US announcements referenced in the plot. JetBlue has a single massive spike on March 13th 2020. Both carriers then spent more than two weeks with a higher level of occurrences of the keyword “cancel” than in February 2020. Spirit Airlines customer service never wrote more than three tweets containing the keyword “cancel” in a day except on March 23rd 2020. Frontier Airlines customer service used the keyword “cancel” only in six tweets over the full month of March 2020.

Based on the observations from the plots in Fig. 5, an important increase in the normalized number of passenger tweets containing the keyword “cancel” can be treated as an unwanted situation that airlines have to deal with.

Definition 1. A keyword-related Twitter situation is defined as an increase over a predefined threshold of the normalized number of passenger-written tweets containing the keyword.

Two metrics to measure the airline reaction to such a situation are proposed here. The aim of the first metric is to measure the effectiveness of the airline response to these keyword-related situations.

Proposed passenger-centric metric 3. The keyword-related Twitter situation quality response score of an airline is the time needed for the airline to bring the normalized number of passenger tweets containing the keyword below a predefined threshold.

The Twitter situation quality response score associated to the keyword “cancel” with a threshold of \( q \) normalized tweets \( n_{c canc}^q \) is calculated using the following formula:

\[
k_{c canc}^q = d_j - d_i
\]

where \( d_0 \) is defined as the first day of the considered period where the normalized number of passenger tweets containing the keyword “cancel” is greater than \( q \), and \( d_f \) is defined as the last day of the considered period where the normalized number of passenger tweets containing the keyword “cancel” is greater than \( q \).

This proposed quality metric measures the time needed for the airline to bring the number of passenger tweets containing the keyword back to a normal state. When measuring the response of long term perturbations, such as the COVID pandemic, this time is measured in days.

The number of passenger tweets containing the keyword is normalized by the total number of passengers carried by the airline over the year 2018.

| Rank | Airline          | Empathy score | Rank | Airline          | Sentiment gap |
|------|------------------|---------------|------|------------------|---------------|
| 1    | Alaska Airlines  | 0.476         | 1    | Frontier Airlines| 0.104         |
| 2    | Southwest Airlines| 0.456         | 2    | Alaska Airlines  | 0.179         |
| 3    | Frontier Airlines | 0.374         | 3    | Delta Air Lines  | 0.228         |
| 4    | Spirit Airlines  | 0.146         | 4    | JetBlue Airways  | 0.228         |
| 5    | United Airlines  | 0.129         | 5    | Spirit Airlines  | 0.237         |
| 6    | JetBlue Airways  | 0.066         | 6    | Southwest Airlines| 0.244         |
| 7    | Delta Air Lines  | 0.029         | 7    | United Airlines  | 0.246         |
| 8    | American Airlines| −0.393        | 8    | American Airlines| 0.260         |

Table 2
Airline ranking based on the proposed empathy score \( \Xi \) and the sentiment gap \( \Delta \) applied to the period of March 1st 2020 to March 31st 2020.
in this case, similarly to the data presented in Fig. 5, and this normalization should be updated with the most recent numbers once they are available.

The aim of the second metric is to measure the communication effort produced by the airline in order to handle the situation linked to the increase of number of tweets containing the keyword under consideration.

Proposed passenger-centric metric 4. The keyword-related Twitter situation quantity response score of an airline is calculated by integrating the number of tweets containing the keyword and written by the airline customer service over the number of days associated to the keyword-related Twitter situation.

The formula used to calculate the Twitter situation quality response score associated to the keyword "cancel" with a threshold of $q$ normalized tweets $\gamma$ is the following:

$$\gamma_{cancel}^q = \int_{d_{cancel}^1}^{d_{cancel}^0} n_{cancel}(t) \, dt$$  \hspace{1cm} (4)$$

where $d_{cancel}^0$ and $d_{cancel}^1$ are the same as for the quality response score $\kappa_{cancel}^q$ in Eq. (3), and $n_{cancel}(t)$ is the number of tweets written by the airline customer service containing the keyword “cancel” on day $t$.

Table 3 presents these two proposed metrics in the case of the keyword “cancel” considering that the predefined threshold indicating when a situation starts and ends is 1. Table 3 illustrates the necessity of considering both the quality response score and the quantity response score hand in hand. Southwest Airlines has the best scores from both perspective but Spirit Airlines has the second best quality response score but the second worst quantity response score. This would indicate that passengers from Spirit Airlines are more resilient to cancellation situations than passengers of the other airlines: They go back to a close-to normal Twitter chatter about cancellation with almost no cancellation related communication efforts on Twitter of Spirit Airlines.

3.2. Refund

Fig. 7 shows the evolution of the normalized number of tweets containing the keyword “refund” and written by passengers normalized by the number of transported passengers per carrier over the year 2018 using BTS data (Bureau of Transportation Statistics, 2020).

Table 3 Airline ranking based on the “cancel”-related Twitter situation quality and quantity response scores applied to the period of March 1st 2020 to April 30th 2020.

| Rank | Airline                  | Quality (days) | Rank | Airline                  | Quantity |
|------|--------------------------|----------------|------|--------------------------|----------|
| 1    | Southwest Airlines       | 11             | 1    | Southwest Airlines       | 50.64    |
| 2    | Spirit Airlines          | 26             | 2    | American Airlines        | 15.34    |
| 3    | United Airlines          | 34             | 3    | Delta Air Lines          | 11.98    |
| 4    | American Airlines        | 35             | 4    | United Airlines          | 11.62    |
| 5    | Delta Air Lines          | 41             | 5    | JetBlue Airways          | 10.28    |
| 6    | Alaska Airlines          | 44             | 6    | Alaska Airlines          | 6.82     |
| 7    | Frontier Airlines        | 53             | 7    | Spirit Airlines          | 0.96     |
| 8    | JetBlue Airways          | 54             | 8    | Frontier Airlines        | 0.11     |
number of occurrences of the keyword “refund” in tweets written by passengers of all four legacy airlines steeply increases at the announcement of Italy’s lockdown and then very slowly decreases. Passengers of Alaska Airlines have an anticipated spike in the number of tweets containing the keyword “refund” at the beginning of March 2020. Fig. 7(b) shows that the increase in the number of tweets containing the keyword “refund” and written by Southwest Airlines passengers is still lower than the number of tweets containing the keyword “refund” and written by the passengers of the other low-cost carriers. The number of tweets containing the keyword “refund” and written by Southwest Airlines passengers gets back to a normal level faster than for the passengers of the other low-cost carriers. The spike in the number of tweets containing the keyword “refund” and written by Spirit Airlines and Frontier Airlines passengers starts only at the announcement of the US travel ban.

Fig. 8 shows the evolution of the number of tweets containing the keyword “refund” and written by airline customer services from February 16th 2020 to May 3rd 2020 for the same eight US airlines.

Fig. 8(a) shows the evolution of the number of tweets containing the keyword “refund” and written by the customer services of the four considered legacy airlines. The initial increase is similar than for the keyword “cancel” (Fig. 6(a)), however there is then a second increase towards the end of March 2020, this increase being most visible within the tweets written by American Airlines customer service. From a low-cost carrier perspective, Fig. 8(b) illustrates the same characteristics as in Fig. 6(b): There are two spikes around the US announcements for the number of tweets containing the keyword “refund” in tweets written by Southwest Airlines customer service, this time with higher fluctuations afterwards, and one major spike on March 13th 2020 for the number of tweets containing the keyword “refund” and written by JetBlue Airways customer service. Only one tweet containing the keyword “refund” was written by Frontier Airlines customer service over the month of March 2020 and none written by Spirit Airlines customer service since February 16th 2020.

The same two metrics associated to the “cancel”-related Twitter situation presented in Section 3.1, i.e. the quality response score and the quantity response score, can be used for this “refund”-related Twitter situation. Table 4 presents these two proposed metrics in the case of the keyword “refund” using the same predefined threshold of 1 for delimiting a Twitter situation.

As for the handling of the “cancel”-related Twitter situation, Southwest Airlines had the most effective (best quality response score) and most proactive (best quantity response score) of the eight airlines. The same resilience is shown by passengers of Spirit Airlines during this “Refund”-related Twitter situation as for the “cancel”-related Twitter situation.

4. Discussion & conclusion

4.1. Score summary

Fig. 9 presents a radar plot for each of the eight considered airlines indicating their normalized scores.

The normalizations were conducted using the following formulas:

\[ \Xi = \frac{1 + \Xi}{2} \]  
\[ \Delta = \frac{1 - \Delta}{2} \]  
\[ \kappa_{\text{keyword}} = 1 - \frac{\kappa_{\text{keyword}}}{\delta T} \]  
\[ \gamma_{\text{keyword}} = \frac{\gamma_{\text{keyword}}}{\delta T} \]

where \( \delta T \) is the number of days of the full period over which the keyword-related Twitter situation response scores are calculated. All-but-one of the normalized scores go from the worst score of 0 to a good score of 1. The score can be greater than 1 in the case of a keyword-related Twitter situation response quantity score, but that scenario did not occur here. Regarding the normalized sentiment gap, a score of 0.5 indicates a normal score of 0, a normalized score of 0 indicates a score of 1 and a normalized score of 1 indicates a score of \(-1\).

4.2. Discussion

As can be seen in Fig. 9, each airline has its own “Twitter profile”. Passengers are then free to integrate these different profiles in their decision

| Rank | Airline          | Quality (days) | Rank | Airline         | Quantity |
|------|------------------|----------------|------|-----------------|----------|
| 1    | Southwest Airlines| 8              | 1    | Southwest Airlines| 32.25    |
| 2    | Spirit Airlines  | 29             | 2    | American Airlines| 22.81    |
| 3    | American Airlines| 31             | 3    | United Airlines  | 7.67     |
| 4    | Alaska Airlines  | 35             | 4    | Delta Air Lines | 7.46     |
| 5    | Delta Air Lines  | 37             | 5    | Alaska Airlines | 4.03     |
| 6    | JetBlue Airways  | 39             | 6    | JetBlue Airways | 3.03     |
| 7    | United Airlines  | 51             | 7    | Frontier Airlines| 0.02     |
| 7    | Frontier Airlines| 51             |      | Spirit Airlines  | 0.00     |
process for choosing the airline that corresponds the most to their travel needs and wants. Traditionally, the airline and airport choices are shown to be based on fare, access time and journey time (Pels et al., 2003; Jung and Yoo, 2014). These studies do not take the airlines reputation among passengers as a decision parameter, and the proposed metrics could provide an additional decision layer for passengers.

For example, some risk-averse passengers could decide to opt for an airline that has better “refund”-related scores if they prefer a refund when flights are cancelled, rather than choosing an airline with a lower fare. Similarly, some passengers can consider that the flight experience is important in their airline decision and use the empathy and sentiment gap scores to help them decide which airline choose. After their experience with the airline, passengers can tweet about it, which will then be taken into account in the next score update. This process corresponds to a feedback loop illustrated in Fig. 10.

On the other hand, airlines can also compare their Twitter profiles provided in Fig. 9 in order to improve their interactions with their passengers. For example, an airline with a clear description of their cancellation procedures on their website could use the “cancel” and “refund” related scores to verify if this information is actually easily accessible to passengers and if adequate communication is made on its availability. For example, a low “cancel” quality score would indicate that passengers already have access to the cancellation information. The proposed metrics can therefore also be used as a part of a feedback loop for airlines, regarding how their policies are implemented and if changes are necessary. Such a feedback loop is illustrated in Fig. 11.

The proposed feedback loops are based on unsolicited feedback, which is usually biased towards extreme negative and extreme positive feedback. The bias potentially remaining in the metrics can be corrected with the collaboration of airlines thanks to the access to their data about their passenger experience surveys.

4.3. Conclusion

The proposed passenger-centric metrics were built using Twitter data, which have the major advantage of being available in real-time, and can therefore be easily updated on an hourly basis if needed. Discussion between federal agencies, airlines and passengers should be undertaken in order to further tune the proposed metrics in order to meet the expectations of all concerned parties.

The proposed metrics have the added benefit of enabling each passenger and airline to actively influence the scores. It should however be emphasized here that the metrics measure essentially the communication quality and quantity between airlines and passengers via Twitter, and should therefore still be complemented with traditional flight-centric measures for completeness.

This study focused on the effects of the travel restriction measures linked to a major disruption taking its course over an important number of days and tailored the proposed metrics for this timespan. Future studies could also investigate into the adaptation of some of these proposed passenger-centric metrics to measure effects on a smaller scale, e.g. over a single day or a few hours.
Srivastava, H., Sankar, R., 2020. Information dissemination from social network for extreme weather scenario. IEEE Trans. Comput. Soc. Syst. 7, 319–328. https://doi.org/10.1109/TCSS.2020.2964253.

Terpstra, T., Stronkman, R., 2012. Towards a Realtime Twitter Analysis During Crises for Operational Crisis Management. p. 10.

Truxal, S., 2020. COVID-19 Airport Slot Rules: What’s Changed and What’s Next for European Airlines?

Tsaur, S.-H., Chang, T.-Y., Yen, C.-H., 2002. The evaluation of airline service quality by fuzzy MCDM. Tour. Manag. 23, 107–115. https://doi.org/10.1016/S0261-5177(01)00050-4.

United States Customs and Border Protection, 2020. Airport wait times. URL: https://awt.cbp.gov.

US Department of State, 2020. Global level 4 health advisory - do not travel. URL: https://travel.state.gov/content/travel/en/traveladvisories/ea/travel-advisory-alert-global-level-4-health-advisory-issn.html.

Van den Heuvel, J., Ton, D., Hermansen, K., 2016. Advances in measuring pedestrians at dutch train stations using bluetooth, wifi and infrared technology. Traffic and Granular Flow ’15. Springer, pp. 11–18.

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. p. 10.

Wan, Y., Gao, Q., 2015. An ensemble sentiment classification system of Twitter data for airline services analysis. 2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, Atlantic City, NJ, USA, pp. 1318–1325 https://doi.org/10.1109/ICDMW.2015.7.

Wang, D., 2007. Methods for Analysis of Passenger Trip Performance in a Complex Networked Transportation System. Doctor of Philosophy. George Mason University, Fairfax, Virginia, USA.

Wang, D., Sherry, D.L., Donohue, D.G., 2006. Passenger trip time metric for air transportation. The 2nd International Conference on Research in Air Transportation.

Watkins, K.E., Ferris, B., Borning, A., Rutherford, G.S., Layton, D., 2011. Where Is My Bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. Transp. Res. A Policy Pract. 45 (8), 839–848. https://doi.org/10.1016/j.tra.2011.06.010.

WorldAtlas, 2020. Which countries are in mandatory lockdown due to COVID-19? URL: https://www.worldatlas.com/which-countries-are-in-mandatory-lockdowns-due-to-covid-19.html

Yu, H.-F., Huang, F.-L., Lin, C.-J., 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. Mach. Learn. 85, 41–75.