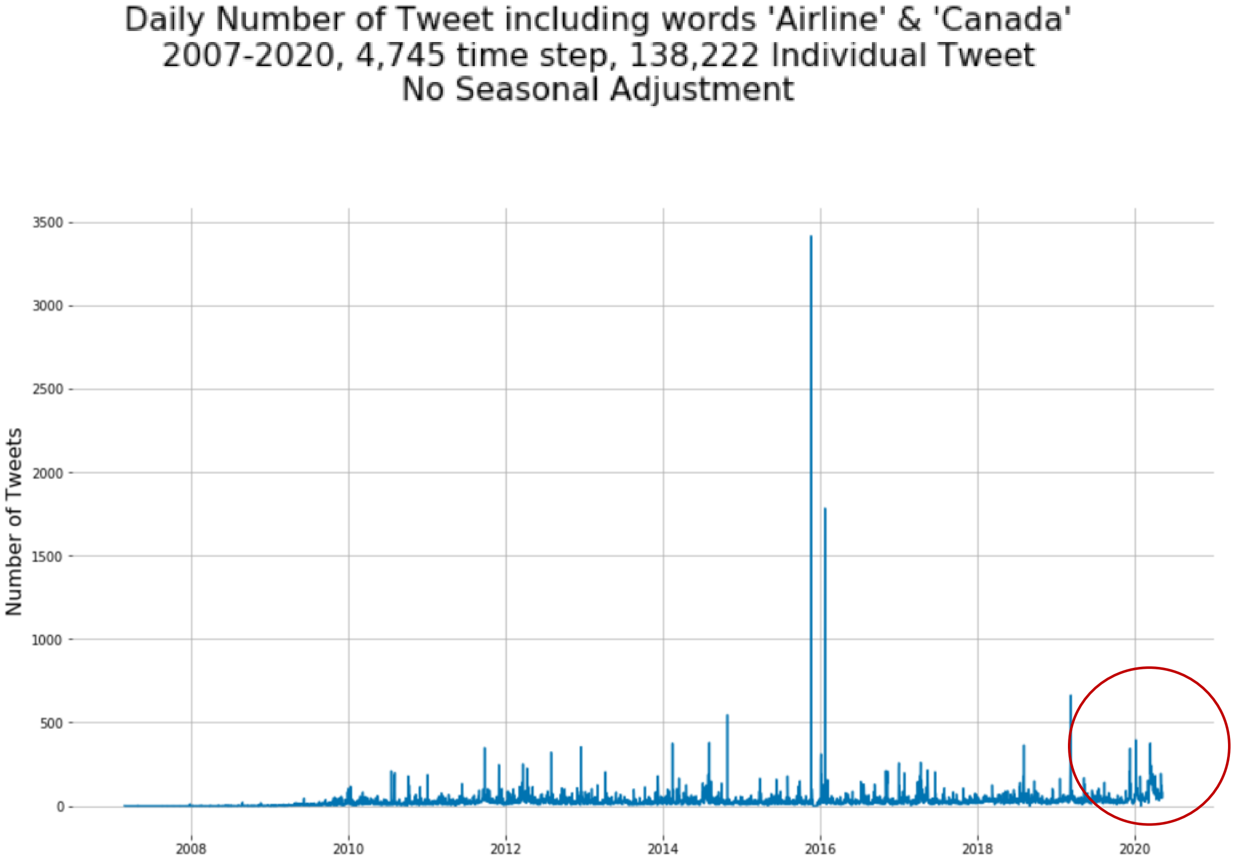
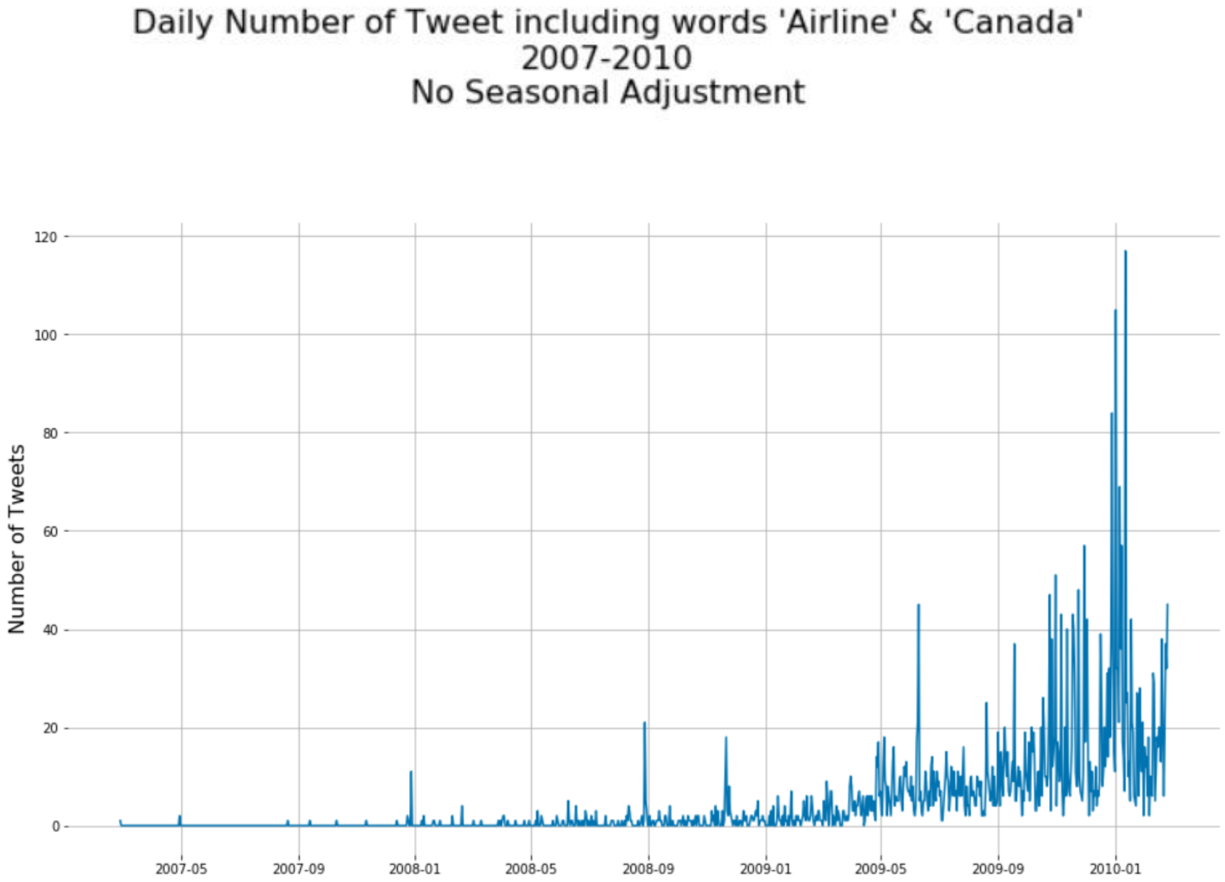
Web Scrapping Transportation Tweets
& Performing Anomaly Detection for Canada

I web scrapped & collected the totality of daily tweets including words ‘*airline’* & ‘*Canada’*. Data range from January 2007 to May 2020. Tweeter was in fact founded in Late 2006. I get a time series of 4,745 daily time step. That include a totality of 138,222 individual Tweet. The dataset is available in our MMA folder. In the dataset, each row is an individual tweet. Variable included are; The date of the tweet, the author of the tweet, the integral text of the tweet, including hashtag if available, & the number of retweet for the tweet. Number of retweet could be used as a way to weight our data point. This approach has not been applied for the moment. I focus purely on the number of tweet published on a daily basis.

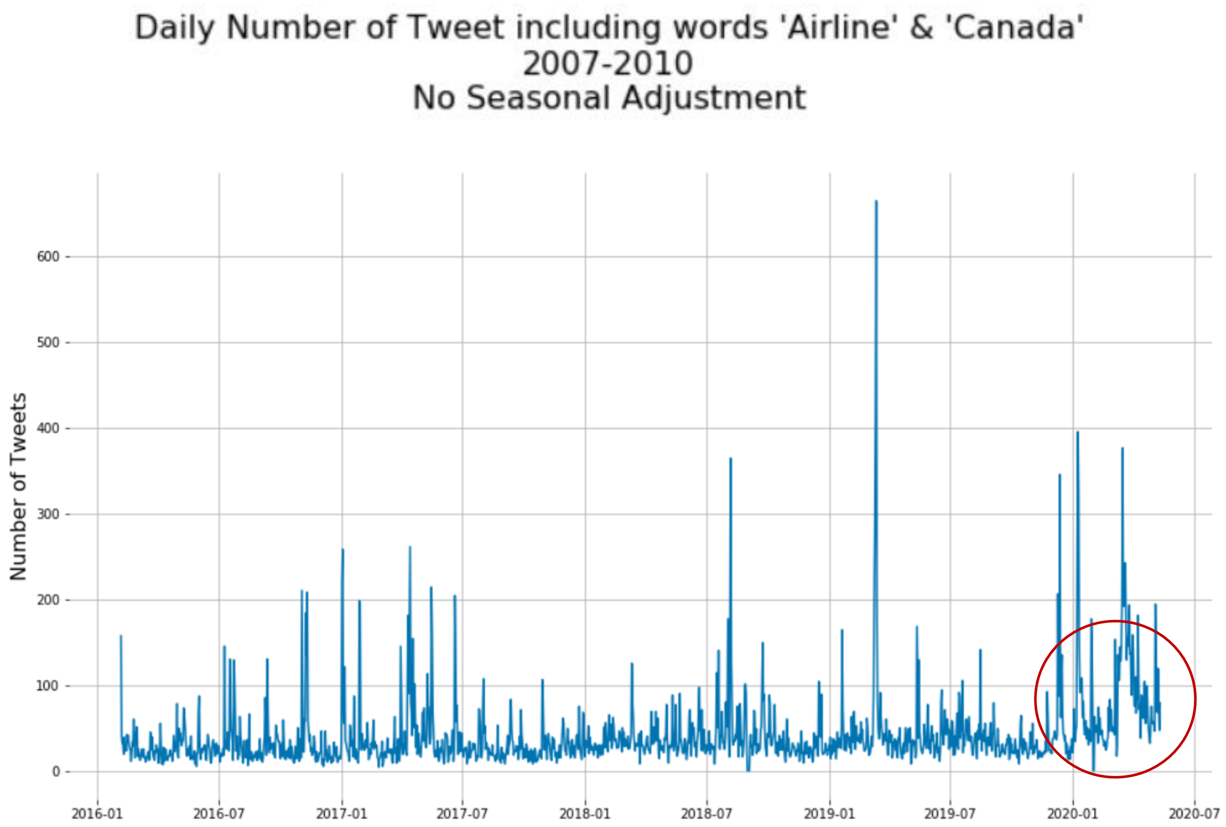
The time series below present the number of tweet, from 2006 to 2020, on a daily basis. There is no seasonal adjustment. The highest peak is in early 2016, but this is not a persistent event. Overall, the series looks quite stationary, in the sense that the historical average & variance seems stable over time. For 2020 though, the pattern seems to change. Let’s look at different segment of the time series.



The time series presented below is the same time series displayed above, but focus on years 2007-2010 only. The series is obviously non-stationary. The level of variance is increasing & the historical average is changing across time. Such patterns in the data were expected. Those are the early days of Tweeter. Users are progressively getting addicted to the platform.



The time series below is the same time series as displayed above, but focus on years 2016-2020 only. Again, the series looks quite stationary overall. In 2016, Tweeter reputation is well established, and user’s daily level of activity is more stable than the early days. On the other hand, the pattern is changing for 2020. The series is becoming less stationary, & more persistent, in the sense that up & down are more progressive, and less drastic. At first sight, we could conclude that such changes are normal, and due to the authors not having enough time to delete their recent tweets & stabilise the series. The analysis below will support the opposite.



The time series below represent a rolling window regression of Augmented Dickey-Fuller test (ADF). It assessed the local level of instability in the data. The x-axis represents the window number. Each window includes 365 days (1 year) of data from the original daily number of tweets series. That is, the first window is for 2007, and the latest window is for the April 2019 – April 2020 interval. The total number of window ADF regression is around 4,500. The ADF test include a constant, with no linear or quadratic trend. For each window, the number of lags of short-run dependant variable are based on best BIC (Bayesian Information Criteria).

The Red Curve represents the ADF statistical value. The more negative (the lower) the value, the more stable the original number of tweet series will be. The 3 Blue Curve represents critical value for 1% (lowest curve), 5% & 10% (highest curve). The green curve represents the p-value of the test. A statistical value above the highest blue curve means the data are not stationary at 10%.

The ADF rolling window regression propose quite a stable time series of information. There is a bit of a noise in the short-run, but overall, there is 3 sub-segment of information in the long-run. Also, the series is quite persistent, in the sense that drastic up & down are present, but rare.

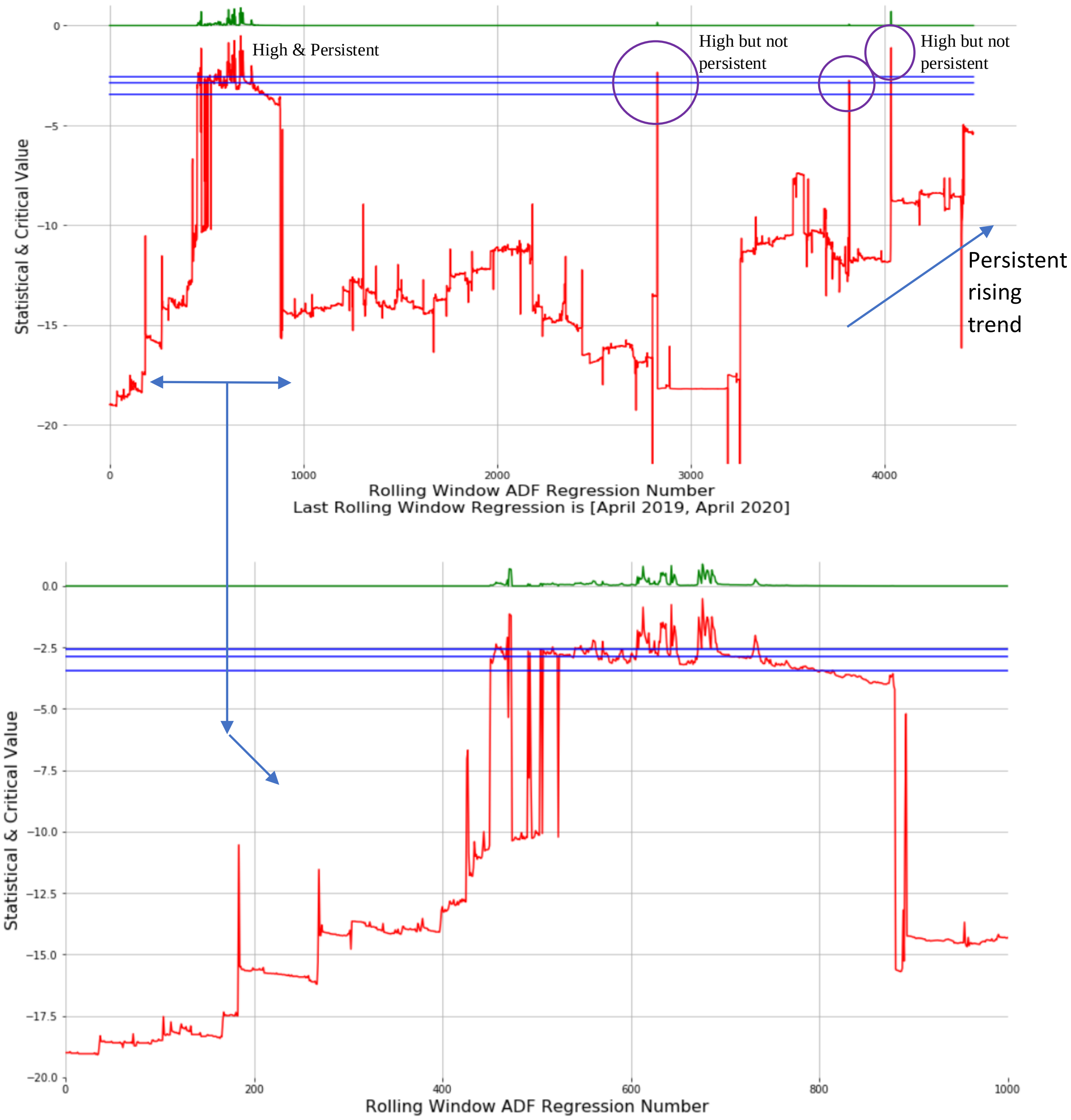
Window 1 to 1,000: Those are the early days of Tweeter. As expected, the level of instability is growing and peak way beyond the critical values. P-values are far above 0%. This segment represent a good reference point. **Window 1,000 to 3,000**: The data are becoming stationary. Tweeter reputation is getting well established. The number of tweet is easier to predict on a daily basis. **Window 3,000 to Present:** The data are still stationary, but relatively less than the second half of the second window. For window 4,000+, we are reaching a 13 years' highest level of instability, but still far from the level of the early days. Again, the level of persistence is there, and cannot be due to users not having enough time deleting recent tweets.

Evolution of the Augmented Dickey-Fuller Test Over Time

ADF Test Based on Daily Number of Tweet including words 'Airline' & 'Canada'
January 2007 - April 2020, 4,745 time step, 138,222 Individual Tweet
No Seasonal Adjustment

365 days (1 year) Rolling Window Regression
With Constant (No Linear or Quadratic Trend)
Max Number of Lags of Short-Run Variables (Delta y) Based on Best BIC Criterion

Red = ADF Statistical Value
Blue = Critical Value (1% (Lowest), 5% & 10% (Highest))
Green = P-Value
Black = Max Number of Lags of Short-Run Variable (Delta y)(None if nb of Lags is 0)



High & Persistent

High but not persistent

High but not persistent

Persistent rising trend

In conclusion, these ADF results are consistent with all observations done on the original time series. Most important of all, the ADF results confirm the persistence of the rising trend of instability during recent years. Monitoring the level of stability for the up-coming months will be important. Comparing these results with the US is essential. Cleaning tweets would be next step.