REFERENCES

[1] Wiener & Bronson. "Facebook's Top Open Data Problems". Facebook Research, 22 Oct. 2014, https://research.fb.com/blog/2014/10/facebook-s-top-open-data-problems/
[2] Pathak. "ETL — Understanding It and Effectively Using It". Medium, 7 Jan. 2019, https://medium.com/hashmapinc/etl-understanding-it-and-effectively-using-it-f827a5b3e54d
[3] Google. "Using Visualizations to Check Your Data". Google. https://www.tensorflow.org/tfx/guide/tfdvusing_visualizations_to_check_your_data

7 APPENDIX

Clean		Dirty	
RowId	0.000000	RowId	0.000000
Source	0.000000	Source	0.000000
Flight	0.000000	Flight	0.000000
ScheduledDeparture	0.000000	ScheduledDeparture	0.476794
ActualDeparture	0.014968	ActualDeparture	0.190283
DepartureGate	0.188292	DepartureGate	1.503791
ScheduledArrival	0.000000	ScheduledArrival	0.464570
ActualArrival	0.025478	ActualArrival	0.197457
ArrivalGate	0.188728	ArrivalGate	1.488697
date	0.000000	date	0.000000

Figure 3: Flights data completeness averages

Clean		Dirty	
RowId	2381.290323	RowId	2381.290323
Source	37.612903	Source	37.612903
Flight	93.483871	Flight	93.483871
ScheduledDeparture	82.935484	ScheduledDeparture	508.064516
ActualDeparture	86.516129	ActualDeparture	662.903226
DepartureGate	67.000000	DepartureGate	147.225806
ScheduledArrival	90.354839	ScheduledArrival	588.225806
ActualArrival	86.322581	ActualArrival	728.419355
ArrivalGate	66.838710	ArrivalGate	146.225806
for_key	82.935484	date	1.000000
date	1.000000		

Figure 4: Flights data uniqueness averages

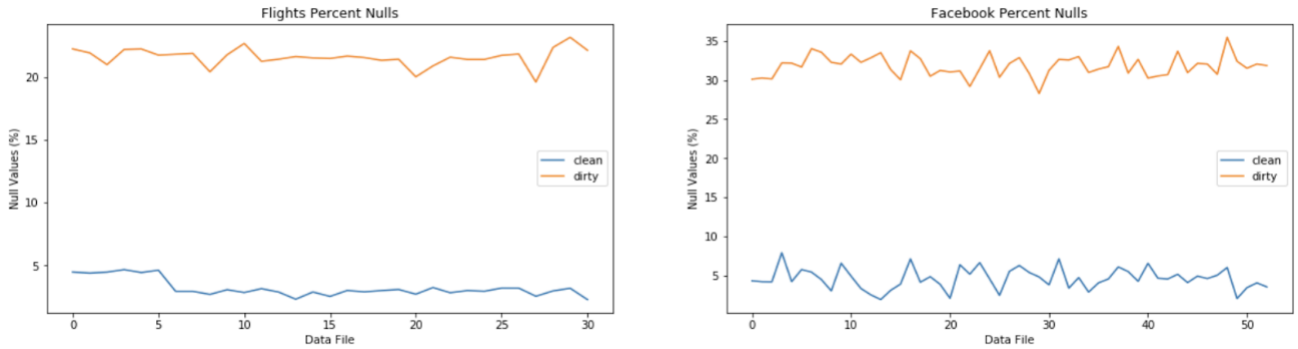


Figure 5: Percentage of null values in each clean and dirty data batch

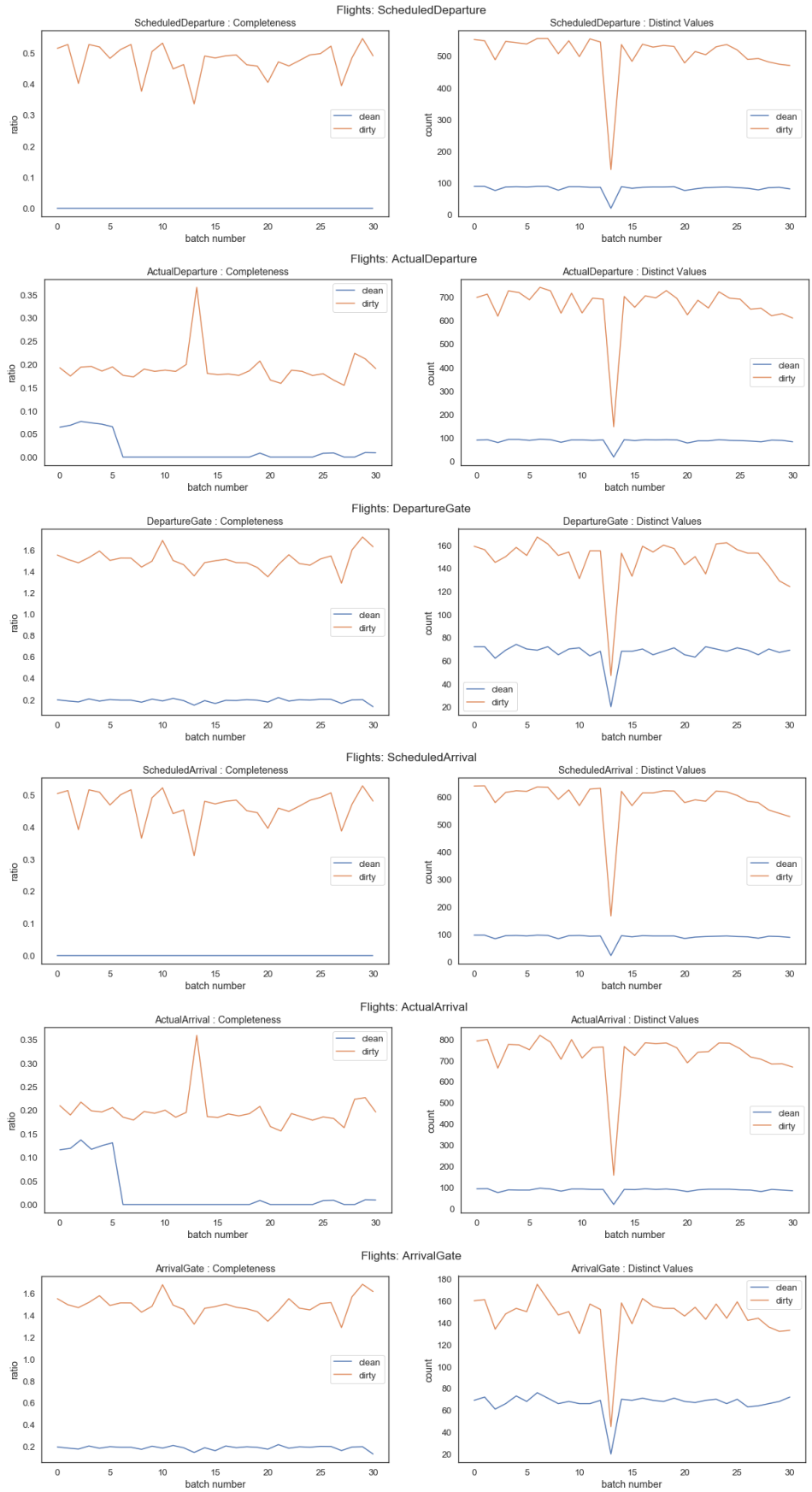


Figure 6: Flights dataset completeness and uniqueness for selected relevant columns

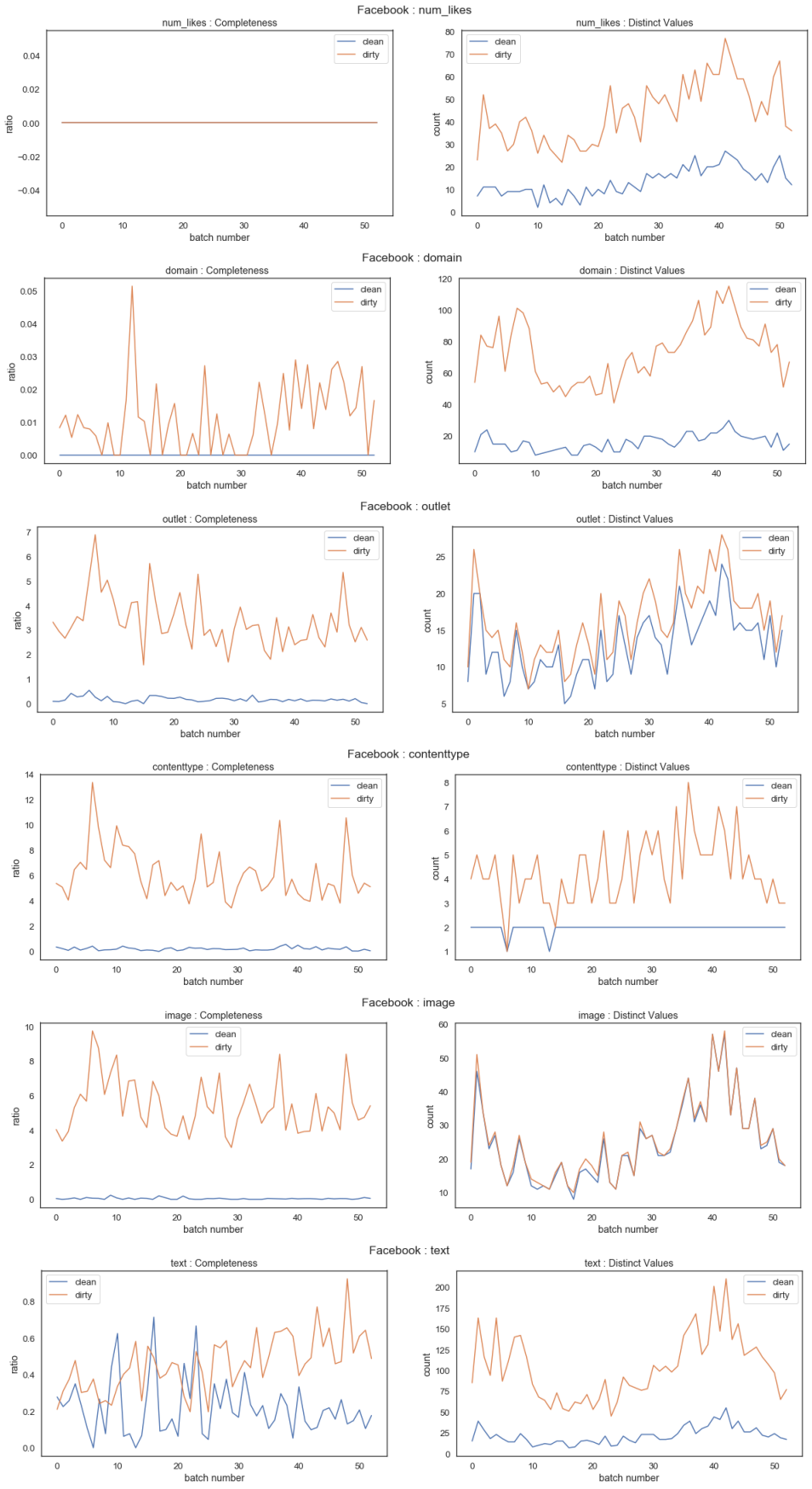


Figure 7: Facebook dataset completeness and uniqueness for selected relevant columns

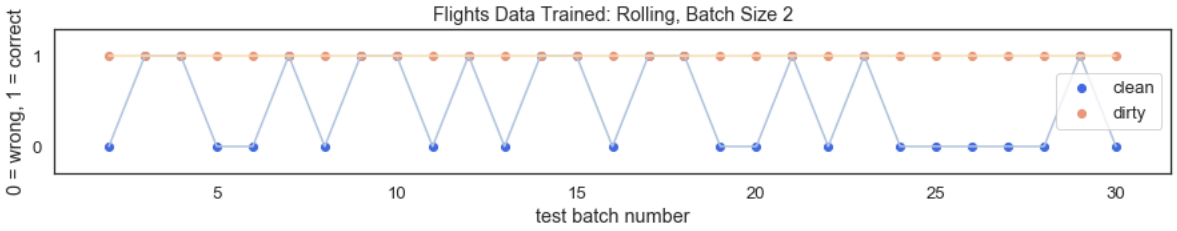
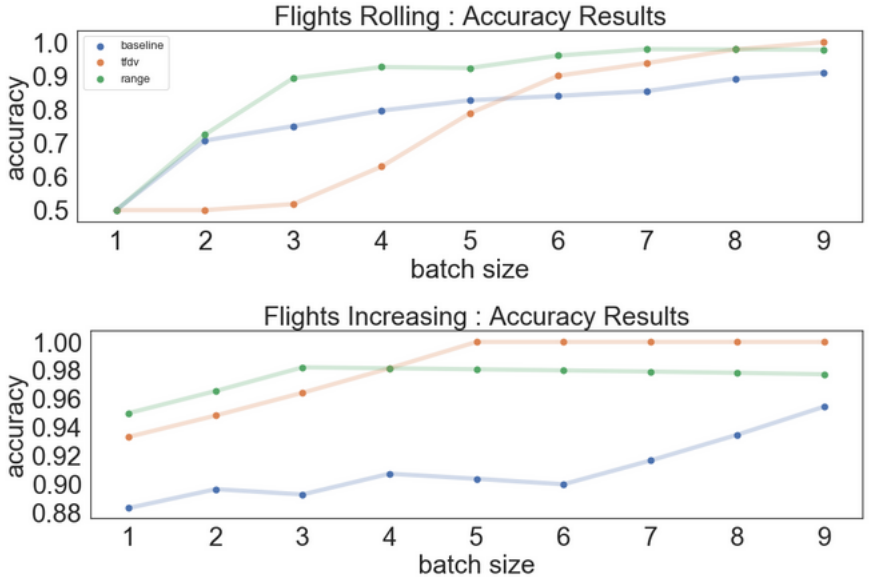


Figure 8: Results for Flights dataset using Rolling training method with batch size 2



si50.si5

Figure 9: Accuracy results for all datasets using both training methods and all criterion methods