Module 22: Harnessing Social Media and Big Data for Transit Business Intelligence

Table of Contents

Module Description

This module provides an overview of social media and big data with a focus on deriving business intelligence for transit operators and planners. This course does not attempt to provide a complete treatment of social media, big data, or business intelligence, but rather focuses on the intersection of the three areas in relation to transit.

1. Introduction/Purpose

Social media platforms have empowered individuals to share their thoughts about all aspects of daily life – from restaurants to movies to political candidates. Increasingly, they are commenting about public transportation services as well. Social channels like Twitter, Facebook, and Instagram have made it easier for transit customers to report on service, maintenance, and safety-related issues, and to do so in real time.

These comments can be a valuable source of business intelligence for transit providers, providing feedback that can inform decisions on operations, planning, and investment. However, despite the widespread use of social media, extracting meaningful information from Facebook posts or Twitter tweets requires specialized analysis techniques.

The volume of social media data is vast. For example, Twitter saw about 500 million posts (or "tweets") per day in 2016. These large datasets require special analytical tools. Even a subset of daily Twitter posts focusing on public transportation is likely to be too large to review with readily available software tools like spreadsheets. Further complicating the analysis challenges, social media posts typically use natural language. This means that analytical approaches must be capable of interpreting noisy and unstructured data.

Big data is the technical approach to managing datasets that are too large or complex for traditional data processing tools. Data mining is a technique used in a big data environment for extracting meaningful or actionable information from large datasets, including those comprising social media posts. Sentiment analysis, sometimes called opinion mining, uses data mining techniques to discern subjective opinions about a topic of interest. Sentiment analysis has been used to examine rider attitudes toward transit service based on a sampling of tweets or other posts.

Business intelligence combines information from multiple sources, including big data, to enable organizations to make data-driven decisions. For transit providers, business intelligence can inform and support service planning, operations, and investment decisions.

2. Samples/Examples

The tables and graphics in the presentation are fully elaborated, and substantive. Because much of the information is visual in nature for the topic in the presentation, as many examples as possible were included in the presentation. If, for some reason, there is a need to shorten the presentation (for example due to time), then those samples and examples removed will be placed in this student supplement.

Many of the examples used can be found on the Internet. The web page references to these examples are listed below:

Web Page Reference
Screen capture from Twitter MBTA CR 01/16/2020 https://twitter.com/MBTA_CR/status/1217795269783379969 Screen capture from Sound Transit Twitter 01/16/2020 https://twitter.com/SoundTransit/status/1217829689848467456
Screen capture from MBTA Transit Police Facebook on 12/17/2019 https://www.facebook.com/MbtaTransitPolice/posts/2566814733404702?__xts__[0]=68.ARC167oHWWQ-LFjC9ytokbuAMrd-HdABN3ApTbKFQYrHdK_dVWZ4YpmBY0Z7IPosKJV83q9QcYVVFUduoDAfGh8FFPClT_ZV3fadpLX3AqvjgdWVa0Hu514MKCN8FDK-u31BaXt_DXfypU9fu5sMmfQiz02-pL8TyfFnZIlEUna9NUbli7334JQYvsOQ1G-d5Co0d4BPAP1kPAY3HRSJZwbKplbcT0LVde6_QXouonieKqltwW8WWOkXSk79f9CsXEkxjvLcYYjo28ewHvQ8bT-Q4HYYU6FgYJjkPVPRUlcbnHo9Ah5q499xXCAE1yzHaPWbcsl5f9763fxpNS-FmwG5Zg&__tn__=-R
Screen capture from LA Metro Instagram 01/16/2020 https://www.instagram.com/p/B7HpVwPh3Fu/
Screen capture from Go transit Twitter 01/16/2020 https://twitter.com/GOtransit/status/1217837017654054912
Screen capture from SEPTA customer service account on 01/16/2020 https://twitter.com/SEPTA_SOCIAL/status/1217951402082930694
Screen capture from Long Beach Transit Instagram 01/16/2020 https://www.instagram.com/p/B7UBwe-Jlui/
Screen capture from MTA Flickr account on 01/07/2020 https://www.flickr.com/photos/mtaphotos/49345936211/
Screen capture from Twitter TTC customer service 01/17/2020 https://twitter.com/TTChelps/status/1218251060315394049
Screen capture from Twitter MBTA account 01/17/2020 https://twitter.com/MBTA/status/1218231659092418565
Screen capture from Twitter TransLink BC 01/17/2020 https://twitter.com/TransLink/status/1218249476223258625
https://trimet.org/about/dashboard/index.htm https://mbtabackontrack.com/performance/#/home http://www.transitheatmap.com/heat.html?austin
https://mbtabackontrack.com/performance/#/detail/reliability/2019-12-01/Subway/green/Green-C/

3. Reference to Other Standards

No standards are referenced in this presentation.

A 2014 Preliminary Report of the ICO/IEC is referenced as a source describing Big Data activities being considered by several standards' organization. (See ISO/IEC JTC1 Information Technology, Big Data Preliminary Report 2014.)

4. Case Studies

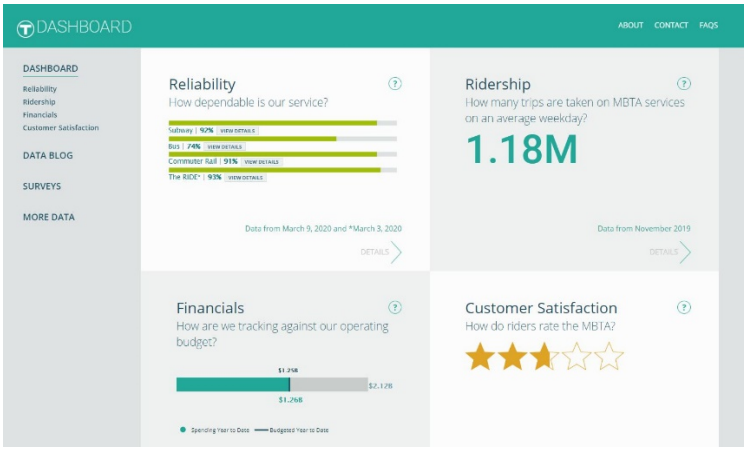
The Massachusetts Bay Transportation Authority (MBTA) has created an online interactive dashboard to provide information on agency performance to internal and external stakeholders. The dashboard was created, in part, to improve transparency and accountability.

The dashboard focuses on four metrics:

- Reliability
- Ridership
- Financials
- Customer satisfaction

Depending on the metric, users can drill down by mode, rail line, peak/off-peak and other characteristics. Users can toggle between visualizations and tabular data. Information is available for several years and updated frequently. Individuals also have the option to download data from the MBTA's data portal for customized analysis.

A companion blog provides detailed discussions about the data sources and metrics used in the dashboard. While the dashboard draws upon multiple data sources, it should be noted that none of the performance indicators are based on social media posts.



The MBTA performance dashboard can be found at https://mbtabackontrack.com/performance/#/home

5. Glossary

To include additional descriptions/acronyms used primarily in the module.

Term	Definition
3 Vs of Big Data	3 salient characteristics of Big Data: variety, volume, and velocity.
Application Programming Interface	An application programming interface (API) is a computing interface exposed by a particular software program, library, operating system or internet service, to allow third parties to use the functionality of that software application. [https://en.wikipedia.org/wiki/Application_programming_interface]
Artificial Intelligence	The study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. [https://en.wikipedia.org/wiki/Artificial_intelligence]
Big Data	Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. [https://en.wikipedia.org/wiki/Big_data]
Business Intelligence	Business intelligence combines information from multiple sources to enable organizations to make data-driven decisions.
Crowdsourced	Crowdsourcing is a sourcing model in which individuals or organizations obtain goods and services, including ideas and finances, from a large, relatively open and often rapidly-evolving group of internet users; it divides work between participants to achieve a cumulative result. [https://en.wikipedia.org/wiki/Crowdsourcing]
Dashboard	A dashboard is a type of graphical user interface which often provides at-a-glance views of key performance indicators (KPIs) relevant to a particular objective or business process. In other usage, "dashboard" is another name for "progress report" or "report." The "dashboard" is often displayed on a web page which is linked to a database that allows the report to be constantly updated. [https://en.wikipedia.org/wiki/Dashboard_(business)]
Data Correlation	In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense, correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. [https://en.wikipedia.org/wiki/Correlation_and_dependence]
Data Lake	A data lake is a concept to describe the storage of data in its raw form. The data lake is a repository of structured and unstructured information.

Term	Definition
Data Pond	A partition of a data lake to limit access, share data resources with another agency. For example, a regional data lake may provide data ponds for separate transit properties.
Data Visualization	Data visualization is the graphical representation of data that communicates insights about data relationships in easy to understand formats.
Key Performance Indicator	A performance indicator or key performance indicator (KPI) is a type of performance measurement.[1] KPIs evaluate the success of an organization or of a particular activity (such as projects, programs, products and other initiatives) in which it engages. [https://en.wikipedia.org/wiki/Performance_indicator]
Machine Learning	Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. [https://en.wikipedia.org/wiki/Machine_learning]
Open Data	Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. [https://en.wikipedia.org/wiki/Open_data]
Open Source Software	Open-source software is a type of computer software in which source code is released under a license in which the copyright holder grants users the rights to study, change, and distribute the software to anyone and for any purpose. Open-source software may be developed in a collaborative public manner. [https://en.wikipedia.org/wiki/Open-source_software]
Performance Measurement	Performance measurement is the process of collecting, analyzing and/or reporting information regarding the performance of an individual, group, organization, system or component. [https://en.wikipedia.org/wiki/Performance_measurement]
Sentiment Analysis	Sentiment analysis, sometimes called opinion mining, uses data mining techniques to discern subjective opinions about a topic of interest.
Similarity Analysis	For the purposes of this ITS PCB Module, similarity analysis is a user-friendly term referring to cosine similarity. Cosine similarity measures the similarity between two vectors of an inner product space (… sometimes referred to as “word space”). It is often used to measure document similarity in text analysis. [https://www.sciencedirect.com/topics/computer-science/cosine-similarity]
Social Media	Social media uses web-based or mobile platforms to enable users to interact with one another.
Stochastic Analysis	Analysis of a stochastic or random process: a mathematical model that usually defines a family of random variables. [https://en.wikipedia.org/wiki/Stochastic_process]

Term	Definition
Topic Maps	As used in this ITE PCB Module 22, refers to the visualization of clusters of unstructured document content based on the results from similarity analysis. For example, one can generate a topic map across a library of thousands of PDF documents to identify which documents contain similar information of interest to the analyst.
Variety	Variety refers to the diversity and inconsistency in the structured and unstructured data present in Big Data.
Velocity	Velocity refers to the speed required to convert input data into output data.
Volume	Volume refers to the quantity of data and growth rate of data.

6. References

- The following are reference materials for those interested in conducting further investigation into the course materials.

- TCRP Report 173: Use of Web-Based Rider Feedback to Improve Public Transit Services. Transit Cooperative Research Program, Transportation Research Board, 2015.
- Best Practices for Transportation Agency Use of Social Media. CRC Press, 2013.
- TCRP Synthesis Report 99: Uses of Social Media in Public Transportation. Transit Cooperative Research Program, Transportation Research Board, 2012.
- National Institute of Science and Technology (NIST) Big Data Working Group Big Data Interoperability Framework (NBDIF), https://bigdatawg.nist.gov/home.php
- American Public Transportation Association (APTA) Leveraging Big Data in the Public Transportation Industry, February 2019, https://www.apta.com/wp-content/uploads/Big-Data-Policy-Brief.pdf
- Bogdan Batrinca and Philip C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms.” AI & Soc (2015).
- Craig Collins, Samiul Hasan, and Satish V. Ukkusuri, “A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data.” Journal of Public Transportation, Vol. 16, No. 2, 2013.
- Stefan Stieglitza, Milad Mirbabaiea, Björn Rossa, and Christoph Neubergerb, “Social media analytics – Challenges in topic discovery, data collection, and data preparation.” International Journal of Information Management 39 (2018).
- ISO/IEC JTC 1 Information Technology, “Big Data Preliminary Report 2014”, 2014

7. Study Questions

The quiz/poll questions and answer choices as presented in the PowerPoint slide are listed below to allow students to either follow along with the recording or refer to the quiz at a later date in the supplement.

Question 1: Which of the following is NOT a source of data business intelligence?

a) Automatic passenger counters (APC)
b) Social media posts
c) Electronic fare collection systems (EFCS)
d) None of the above

Question 2: Which of the following is NOT a source of social media data for business intelligence?

a) Agency marketing posts
b) Customer complaints
c) Customer questions
d) Peer-to-peer communications

Question 3: Which of the below is not one of the 3 V characteristics of big data?

a) Velocity
b) Viscosity
c) Variety
d) Volume

Question 4: Which of the below is not a step described in Big Data processing?

a) Data preparation
b) Data field quantization
c) Data analysis
d) Data acquisition

Question 5: Agencies of researchers have used social media to determine which of the following?

a) Where to upgrade bus shelters
b) How to understand rider sentiment
c) Where to add fare enforcement
d) How to report non-fare revenues

8. Icon Guide

The following icons are used throughout the module to visually indicate the corresponding learning concept listed out below, and/or to highlight a specific point in the training material.

1) **Background information:** General knowledge that is available elsewhere and is outside the module being presented. This will be used primarily in the beginning of slide set when reviewing information readers are expected to already know.

2) **Tools/Applications:** An industry-specific item a person would use to accomplish a specific task and applying that tool to fit your need.

3) **Remember:** Used when referencing something already discussed in the module that is necessary to recount.

4) **Refer to Student Supplement:** Items or information that are further explained/detailed in the Student Supplement.

   Example: Additional information on a standard, additional case studies or examples that don't fit into the PowerPoint itself, external resources, etc.

5) **Example:** Can be real-world (case study), hypothetical, a sample of a table, etc.

6) **Checklist:** Use to indicate a process that is being laid out sequentially.