Reading Between The Lines: Text-Based Analytics in the Services Sector

Anand Gopal
University of Maryland
The Services Sector
Academic Research in Services
However, new opportunities...

- Combination of “big data”, unstructured data, and tractable analytical methods
- Allows the re-examination of problems but with enhanced visibility and precision

  - “These data, sometimes available from public sources but other times obtained through data-sharing agreements with private firms, can help to create more granular and real-time measurement of aggregate economic statistics. The data also offer researchers a look inside the “black box” of firms and markets by providing meaningful statistics on economic behavior such as search and information gathering, communication, decision-making, and micro-level transactions.”, Einav and Levin 2014

  - “As the supervised machine learning research community and other disciplines continue to join together in pursuit of solutions to real-world policy problems using big data, we expect that there will be even greater opportunities for methodological advances, as well as successful implementations, of data-driven policy.”, Athey 2017
What kind of data are we talking about?

**TERMINOLOGY**

**SOME APIs**
Data that has a standard Web service

**NO APIs**
Data that has no standard Web service and requires alternative methods of integration

**INTERNAL**
Data that resides behind an organization’s firewall

**EXTERNAL**
Data that resides outside of an organization’s firewall

**UNSTRUCTURED**
Data that does not have a pre-defined data model or is not organized in a pre-defined manner

**STRUCTURED**
Data that resides in a fixed field within a record or file

**VELOCITY**
The rate at which data is generated and changed

**VARIETY**
The number of different data sources and types

**VOLUME**
The average quantity of data units per category

---

**ARCHIVES**
Archives of scanned documents, statements, insurance forms, medical record and customer correspondence, paper archives, and print stream files that contain original systems of record between organizations and their customers

**DOCS**
XLS, PDF, CSY, email, Word, PPT, HTML, HTML 5, plain text, XML, JSON, etc.

**MEDIA**
Images, videos, audio, Flash, live streams, podcasts, etc.

**DATA STORAGE**
SQL, NoSQL, Hadoop, doc repository, file systems, etc.

**BUSINESS APPS**
Project management, marketing automation, productivity, CRM, ERP content management systems, HR, storage, talent management, procurement, expense management, Google Docs, intranets, portals, etc.

**PUBLIC WEB**
Government, weather, competitive, traffic, regulatory, compliance, health care services, economic, census, public finance, stock, OSINT, the World Bank, SEC/Edgar, Wikipedia, IMDb, and other Web services

**SOCIAL MEDIA**
Twitter, LinkedIn, Facebook, Tumblr, Blog, SlideShare, YouTube, Google+, Instagram, Flickr, Pinterest, Vimeo, Wordpress, IM, RSS, Review, Chatter, Jive, Yammer, etc.

**MACHINE LOG DATA**
Event logs, server data, application logs, business process logs, audit logs, call detail records (CDRs), mobile location, mobile app usage, clickstream data, etc.

**SENSOR DATA**
Medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines, office buildings, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery, etc.
Unstructured data – especially text

Unstructured data poses several challenges:

- High dimensionality, predictive accuracy versus causal inference, program or treatment evaluation, supervised versus unsupervised ML, interpretability, newer statistical tools (Kleinberg et al. 2015, Athey 2017)
Example: Using Text Analytics to Address A Policy Problem - *Hygiene*

– Setting: The **Restaurant Sector**
  * ~$800B in sales
  * 14.7 million employees (10% of US workforce)
  * Half of all adults have worked in the restaurant industry at some point (National Restaurant Association)

*Watch Where You Eat: On the Use of Online Reviews in Identifying Moral Hazard in New York City Restaurant Hygiene Inspections*

Jorge Mejia\(^{a,1}\), Shawn Mankad\(^{b}\), and Anandasivam Gopal\(^{c}\)

\(^a\)Indiana University; \(^b\)Cornell University; \(^c\)University of Maryland
The Birth of a New Pastry: The Cronut®
Then someone on social media reported...
Cronut Bakery Dominique Ansel Shut Down For "Severe Mouse Infestation"

BY JEN CARLSON IN FOOD ON APR 4, 2014 3:39 PM

Dominique Ansel Bakery, home of the Cronut, shut down by health officials for mice
The bakery has received an “A” health grade since opening in 2011, the rep said.

Health officials have shut down Ansel's Spring.
So do hygiene inspections work?

- They work…
  - Reduce rates of foodborne illness by 13-15% (Bucholz et al. 2002; Irwin et al. 1989)
  - Public grades of restaurants are even more effective (Jin and Lee 2014; Jin and Leslie 2013; Simon et al. 2005)

- …but grade inflation reigns (Ho 2012)
  - System shifts resources from critical hazards to resolving grade disputes

- … and they can be potentially worked around
  - Inspections are not “real-time”, discrete events
  - Classic case of information asymmetry
Moral Hazard

– High-quality products are more expensive to produce than low-quality and buyers cannot observe quality perfectly (Holmstrom 1979; Harris and Raviv 1979; Grossman and Hart 1983; Shapiro 1986)

– Moral hazard – protection from risk changes agent behavior, leaving other entities to bear costs of non-compliance

– How do you address moral hazard?
  • Certification
  • Information system – provide observability
Moral hazard in hygiene inspections?

- Restaurants are required to “clean up” so as to maintain hygiene standards
- “Surprise” hygiene inspections lead to hygiene certification
  - Next hygiene inspection – 12-18 months
- Will restaurants retain hygiene practices post-certification?
  - Costly and cumbersome
  - Not always directly visible by consumers
  - Regulatory burden – “over-reach”
What do we propose here?

– Ensuring hygiene quality between inspections → reducing moral hazard
  • “information system”, “real-time” observability
– One potential source for “real-time”: social media

Machine learning in economics (Athey 2015; Athey and Imbens 2015)

Public health surveillance using social media (Harrison et al. 2014; Kang et al. 2013)
Two Primary Objectives

1. Can social media be used to *monitor* the hygiene of restaurants?

2. Is there *evidence* of moral hazard in the NYC hygiene inspection program?
NYC Hygiene Inspection Program

• Inspection Process
  1. Initial inspection ("A" or "pending")
  2. If "pending", re-inspection 1-3 months later
  3. Next inspection ~one year later

• Critical & non-critical violations

Score of 0-13  Score of 14-27  Score ≥ 28
The Program is a “Success”

- 30% “As” in 2010, 80% in 2014 (DOHM 2014)
- New Yorkers like the program (CUNY 2010, 2011, 2012)

91% Approve of public grades
89% Consider grades when dining
77% Strongly prefer A-grade restaurants

- Credited with decrease in foodborne illness
  - 2.1M cases in 2009, 1M in 2014
  - Salmonella ↓14% 2010 to 2013
90% of restaurants currently “A” grade

Percent of Restaurants with A Grades

Bronx
Brooklyn
Manhattan
Queens
Staten Island
New York City

2011 2012 2013 2014 2015

0 10 20 30 40 50 60 70 80 90 100
Two paths to get an “A”

Path 1:

Initial Inspection

Path 2:

Re-Inspection
Inspection Scores by Path to “A”
Moral Hazard In Our Context

- Initial Inspection 1: P-A-P-A 21.93
- Re-Inspection 1: A-A 9.45
- Initial Inspection 2: P-A-P-A 8.75
- Re-Inspection 2: A-A 8.71
- Group:
  - ○ P-A-P-A
  - ● A-A
Strategy: use *online reviews* of restaurants from social media.

Social Media
Structured & unstructured data

SMASH
Social MediA Sourced Hygiene Dictionary

“Live” Restaurant Hygiene Scores

Food icons with different colors and scores.
Social Media Sourced Hygiene Dictionary (SMaSH)

• Dictionary creation (Tsai et al. 2013; Feldman 2013; Liu 2015)
  – Step 1: Create initial seed list of words
  – Step 2: Augment with synonyms using WordNet
  – Step 3: Manually curate final word list

Naïve Bayes classifier (Liu 2015)
Building the List of Seed Words

- Sample 1200 reviews for training (see next slide)
- Use MTurk to label review $j$ as hygiene-related or not
- Compute $P(w_{jk}|d_j = 1)$, where $w_{jk}$ is the number of times word $k$ appears in review $j$
- Keep top 5% of $P(w_{jk}|d_j = 1)$ as seed list
Training Data for Seed List

• Yelp reviews of NYC restaurants
  – 21,000 restaurants
  – 1.3 M reviews from 2004 to 2015
  – 85% of restaurants matched to inspections

• Training data for seed list
  – Identify restaurants with worst inspection scores
  – Randomly sample 1,200 negative reviews from before start of program in 2010
Using the *Crowd* to Label Reviews

• 1,200 MTurk subjects (Goodman et al. 2013)
  – Procedure: Assign each subject to one review
  – Measure: Review related to hygiene? (7-point scale adopted from Egan et al. (2006))
  – Manipulation: Positive or negative review

• Results: 15% of reviews hygiene related
SMaSH Score

Number of dictionary words in document $j$:

$$WC_j = \sum_k w_{jk} \delta_k$$

- $w_{jk}$: number of times word $k$ appears in review $j$
- $\delta_k$: 1 if word $k$ included in SMaSH and 0 otherwise

Summarize by restaurant or time period and use in econometric model
Can SMaSH *explain* inspection scores?

**Longitudinal Mixed Effects Model**

\[
INS_{it} = \beta_1 i + \beta_2 WC_{it} + \beta_3 Rating_{it} + \\
\beta_4 Reviews_{it} + \beta_4 Price_i + \gamma Chars_i + \varepsilon_{it}.
\]

- \(INS_{it}\) = Inspection score for restaurant \(i\) during time period \(t\) (quarter)
- \(WC_{it}\) = SMaSH score for restaurant \(i\) during time period \(t\)
- \(Chars_i\) = Other characteristics of restaurant \(i\) (e.g. parking, payment method, ambiance)

- Significant model: each word count \(\rightarrow 0.5\) point increase in inspection score \((p<0.01)\)
What happens to SMaSH after inspections?

• PAPA group shows familiar patterns…

• As do the AA group
Do restaurants retain their hygiene habits post-rating?

- If SMaSH acts as an “information systems”, we predict inspection score 90 days out as a function of SMaSH word counts.
  - Include restaurant and reviewer fixed effects

Longitudinal Mixed Effects Model

\[ WC_{it} = \beta_1 + \beta_2 Offset_{it} + \beta_3 Offset_{it} AA_i + \beta_4 Offset_{it} PAPA_i + \beta_5 Rating_{it} + \beta_6 Reviews_{it} + \beta_7 Price_i + \gamma OtherChars + \varepsilon_{it} \]
PAPA restaurants clearly regress...

Most AA restaurants likely to retain their hygiene ratings

Roughly 20% show predicted ‘C’ levels, while 40% are in the ‘B’ range
Implications for Policy

– Restaurant inspections – successful, but potentially overstating the extent of success
  • Can online reviews help identify “repeat” offenders?
  • Help in directing inspector effort optimally

– Machine learning + economic testing
  • Online reviews in the public health context
  • Trust in the “crowd”
The Promise of Text

- “Expect the importance of text as data in economic analysis to grow”, Gentzkow et al. (2017)

• Acknowledgements: Kislaya Prasad, Ritu Agarwal, Jen Golbeck, Michael Trusov