Incorporating Text Data in Predictive Analytics: An Application Using Automobile Complaint and Defect Data

Presented at
CAS Cutting Edge Tools for Pricing and Underwriting Seminar
October 3, 2011 (Baltimore, MD)

Presented by
Philip S. Borba, Ph.D.
Milliman, Inc.
New York, NY

Casualty Actuarial Society -- Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

OVERVIEW OF PRESENTATION

1) General Types of Data in Property-Casualty Claim Files

2) Examples of "Real World" Unstructured Data
   - USDOL: Fatality and Catastrophe Injury Data File
   - NHTSA: Complaint Data

3) Processing Unstructured Data

4) Incorporating Unstructured Data into Data Analytics

Strong caveat: Statistics in this presentation are for a very limited number of narrowly-defined cases from USDOL and NHTSA public-access databases. The cases and statistics are intended to demonstrate the principles of processing and analyzing unstructured data, and not for drawing conclusions or inferences concerning the subject matter of the data.

(1) General Types of Data

CLAIM MASTER FILE
("structured data")

Formats:
- one record per claim/claimant

Typical Fields:
Claim_Number
Claimant_Number
Line_of_Business / Coverage
Date_of_Loss
Date_Reported
Date_Closed
Total_Incurred_Loss
Total_Paid_Loss
Total_Recovery
Total_Adj_Expenses
Case_Narrative (special case)

TRANSACTION DATA

Types of Transactions:
- payments
- reserves

Formats:
- one record per trans (multiple records per claim or claimant)

Typical Fields:
Claim_Number
Claimant_Number
Line_of_Business / Coverage
Date_of_Transaction
Type_of_Transaction (codes)
Amount ($)

ADJUSTER NOTES
("unstructured" data)

Free-form text fields

Types of Text Information:
- diary entries
- adjuster notes
- system-generated information

Formats:
- one record per adjuster note
- one record with all adj notes for a single claim, with delimiters

Typical Fields:
Claim_Number
Date_of_Entry
Adjuster_Name
Type_of_Note
Adjuster_Note

Why Unstructured Data?

- Why the interest in unstructured data?
  - Claim segmentation
    - Open claims can be segmented for claim closure strategies (eg., “waiting for attorney response,” “waiting for IME”)
    - Improved claim triage, especially during times of high volume (eg., disasters)
    - Improved recognition of claims with attorney representation
  - Predictive analytics
    - Able to capture information not available in structured data
- Types of unstructured data
  - Claim adjuster notes
  - Diary notes
  - Underwriting notes
  - Policy reports
  - Depositions

2) EXAMPLES OF "REAL WORLD" UNSTRUCTURED DATA

- US Department of Labor
  - Fatality and Catastrophe Investigation Summary
    - Accessible case files on completed investigations of fatality and catastrophic injuries occurring between 1984 and 2007

- National Highway Traffic Safety Administration
  - Four downloadable files
    - Complaints
    - Defects
    - Recalls
    - Technical Service Bulletins

USDOL Fatality and Catastrophe Injury File -- Characteristics

- Cases are incidents where OSHA conducted an investigation in response to a fatality or catastrophe. Summaries are intended to provide a description of the incident, including causal factors.
- Public-access database has completed investigations from 1984 to 2007.
- 15 data fields
  - Structured data fields
    - Date of incidence, date case opened
    - SIC, establishment name
    - Age, sex
    - Degree of injury, nature of injury
  - Unstructured data fields
    - Case summary (usually 10 words or less)
    - Case description (up to approximately 300 words)
    - Key words (usually 1 to 5 one-word and two-word phrases)

USDOL: Sample Case -- Fatality

- Accident: 202341749
- Event Date: 01/23/2007
- Open Date: 01/23/2007
- SIC: 3731
- Degree: fatality
- Nature: bruise/contusion/abrasion
- Occupation: welders and cutters
- Case Summary: Employee Is Killed In Fall From Ladder
- Employee #1 was a welder temporarily brought in to assist in a tanker conversion. Employee #1 was using an arc welder to attach deck angle iron. Periodically Employee #1 had to adjust the resistance knobs. According to the only witness, Employee #1 stepped off the ladder and held onto metal angle iron (2.5 ft apart) to allow the witness to pass. Employee #1 apparently slipped and fell approximately 20 foot to his death.
- Keywords: slip, fall, ladder, welder, arc welding, contusion, abrasion

USDOL: Sample Cases

- Dates of injury: 2006/2007
- SIC: 37
- 120 cases
  - 55 fatalities (46%)
  - 65 catastrophic injuries (54%)

- Present interest
  - Can case descriptions be used to segment claims into fatality/non-fatality cohorts?

NHTSA Downloadable Data Files

- Complaints: defect complaints received by NHTSA since Jan 1, 1995.
- Defect Investigations: NHTSA defect investigations opened since 1972.
- Recalls: NHTSA defect and compliance campaigns since 1967.
- Technical Service Bulletins: Manufacturer technical notices received by NHTSA since January 1, 1995.

NHTSA Complaint File

- Complaints are vehicular related, including accessories (eg, child safety seats)
- Over 825,000 records
- Approximately 620,000 records with a VIN number
- 47 data fields
  - Manufacturer name, make, model, year
  - Date of incident
  - Crash, fire, police report
  - Component description (128 bytes)
  - Complaint description (2,048 bytes)

NHTSA Complaint File – Sample Case 1

- Number of injuries: 0
- Number of deaths: 0
- Police Report: N
- Component description: service brakes, hydraulic: foundation components

- Complaint: “brakes failed due to battery malfunctioning when too much power was drawn from battery for radio”

NHTSA Complaint File – Sample Case 2

- Number of injuries: 1
- Number of deaths: 0
- Police report: Y
- Component description: air bags: frontal

- Complaint: Accident. 2008 Mercedes c-350 rear ended a delivery truck. Mercedes began smoking immediately and caught fire within one minute. Within 3-5 minutes engine compartment and passenger compartment were fully engulfed in flame. Driver escaped before car burned. Airbags deployed in this front end crash. Driver had concussion and facial injuries from hitting, possibly steering wheel. Driver sustained other injuries as well.

NHTSA: Sample Cases

- Model year: 2008
- Complaints with a VIN
- 4,478 cases
  - 6% with casualty

    ("casualty" defined to be a complaint with an injury or death)

- Present interest
  - Can case descriptions be used to improve the ability to predict the incidence of a casualty?

OVERVIEW OF PRESENTATION

1) General Types of Data in Property-Casualty Claim Files

2) Examples of “Real World” Unstructured Data
   - USDOL: Fatality and Catastrophe Injury Data File
   - NHTSA: Complaint Data

3) Processing Unstructured Data

4) Incorporating Unstructured Data into Data Analytics

(3) PROCESSING UNSTRUCTURED DATA

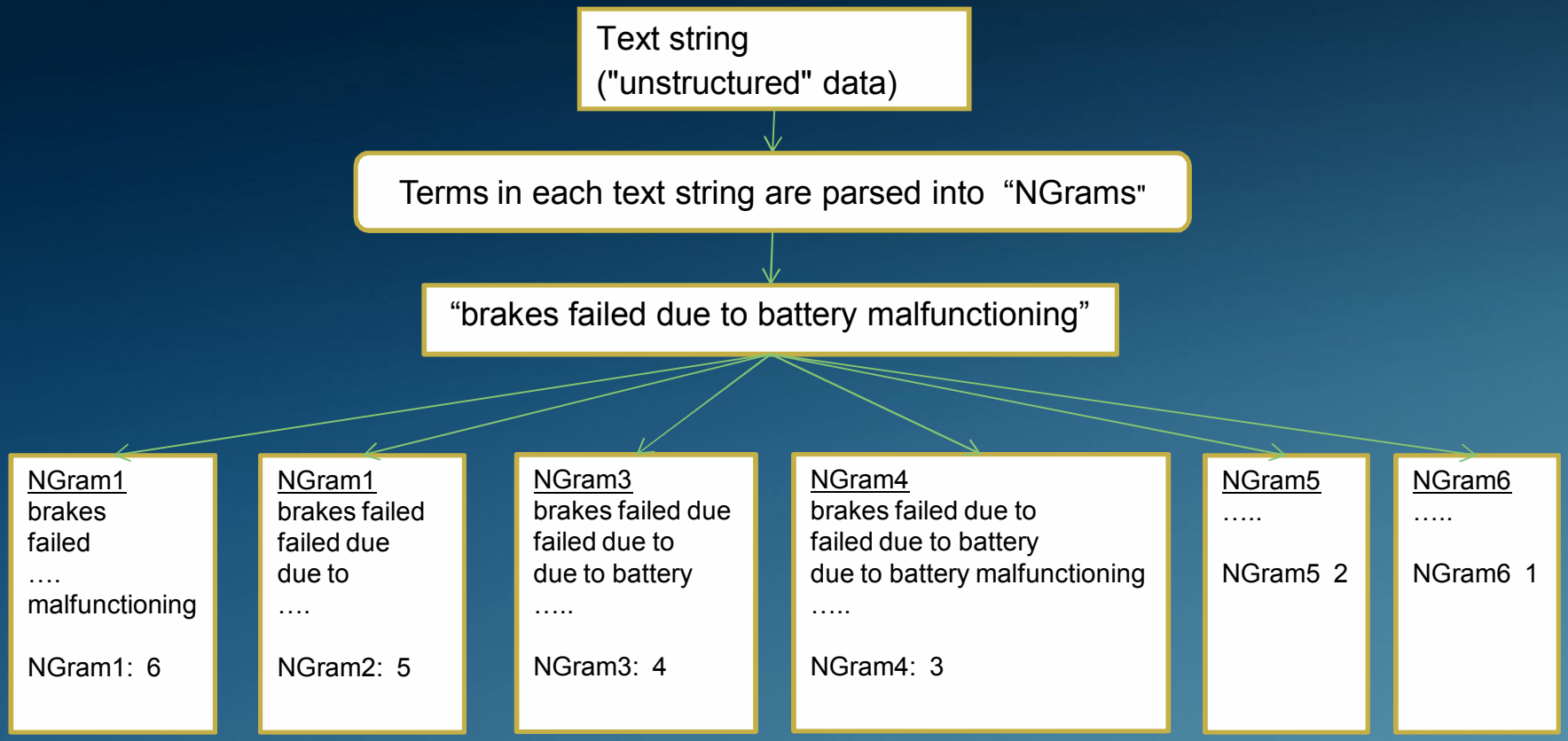
- Parsing Text Data Into NGrams
- Number of NGrams Created from USDOL and NHTSA Sample Cases
- Ngram-Flag Assignments
- Examples of Ngram-Flag Assignments using NHTSA Data

Summary Characteristics of USDOL and NHTSA Sample Cases

- Number of cases and number of terms in sample cases

	USDOL	NHTSA
Number of cases	120	4,478
Number of bytes in case descriptions		
Average number of bytes	531	1,103
Median number of bytes	428	689
Q1 / Q3 number of bytes	275 / 691	418 / 1,284
Maximum number of bytes	1,935	19,383

Parsing Text Data Notes Into NGrams



Number of NGrams Created from USDOL and NHTSA Sample Cases

- Each case description is parsed into NGram1-NGram6
- Process removes certain NGram1-NGram3 not expected to be needed in any claim-segmentation or analytics
- Longer case descriptions increase the number of NGrams per case

	USDOL	NHTSA
Number of cases	120	4,478
Size of NGram		
NGram1	8,468	682,234
NGram2	10,722	847,523
NGram3	10,729	842,700
NGram4	10,606	832,777
NGram5	10,484	821,382
NGram6	10,363	810,139

USDOL: Summary Stats from NGram Creation

- In sample cases, certain Ngrams were more (less) prevalent among fatal injuries.

Ngram	Percent of Cases where Case was a Fatality
Ladder	75%
Forklift	70%
Conveyor	67%
AVERAGE	**46%**
Was Hospitalized	3%
Amputation	0%

- Strong caveat: Statistics in this presentation are for a very limited number of narrowly-defined cases from USDOL and NHTSA public-access databases. The cases and statistics are intended to demonstrate the principles of processing and analyzing unstructured data and not for drawing conclusions or inferences concerning the subject matter of the data.

NGrams-Flag Assignments

- Flags identify NGrams with similar concepts

- For NHTSA Complaint data:
  - Acceleration Sudden
  - Air Bag
  - Another Vehicle
  - Brake Failure
  - Manufacturer Defect
  - Safety Issue

- Challenges
  - Design: concepts captured in flags can be very specific (using few, narrowly specified Ngrams) or very broad
  - Operational: misspellings, punctuation, synonyms, special abbreviations by data source

Examples of Ngram-Flag Assignments for NHTSA Data

Ngram	Flag
Air bag	Air Bag
Air bags	
Airbag	
Airbags	
Brakes defective	Brake Failure
Brakes failed	
Brake failure	
Defective brakes	
Accelerated suddenly	Sudden acceleration
Sudden acceleration	

OVERVIEW OF PRESENTATION

1) General Types of Data in Property-Casualty Claim Files

2) Examples of "Real World" Unstructured Data
   - USDOL: Fatality and Catastrophe Injury Data File
   - NHTSA: Complaint Data

3) Processing Unstructured Data

4) Incorporating Unstructured Data into Data Analytics

4) INCORPORATING UNSTRUCTURED DATA INTO DATA ANALYTICS

- Analytics File
- Flags of Interest for Analytics of NHTSA Data
- Types of Analytics
- Development of a Probability Model

Analytics File

- Analytics File
  - Combines structured data and flags created from the unstructured data
  - One record for each complaint (or claim, claimant, injury, etc.)

- NHTSA example
  - Developed from structured data
    - Manufacturer name, make, model, year
    - Date of incident
    - Crash, fire, police report
  - Developed from unstructured data (case descriptions)
    - 0/1 flags for brake failure, accelerator failure, sudden acceleration

Flags of Interest for Analytics of NHTSA Data

- Flags developed using NHTSA data
  - Accelerator Sudden
  - Air Bag
  - Another Vehicle
  - Brake Failure
  - Brake Pedal
  - Engine Failure
  - Fuel System
  - Manufacturer Defect
  - Safety Issue
  - Steering Failure
  - While Driving
- Present flags are for demonstration. The number of flags is limited only by imagination and practical considerations.

Types of Analytics

- Summary Statistics
  - Frequencies and averages for different breakdowns (eg, state, line of business, 0/1 flags, yes/no for outcome measure)
- "Story" Scenarios (Predictive Modeling I)
  - A "Story" is a particular set of results for a given set of flags.
  - For example, for a given set of four flags (eg, brake failure, acceleration failure, safety hazard, steering failure), the target result is a series of four "0" and "1" flags.
  - Claims fitting a Story are flagged for the target outcome (eg, likely casualty).
- Probability Model (Predictive Modeling II)
  - Model can produce an estimate for the probability of an outcome.
  - Number of factors and weights assigned to the factors can vary.

Development of a Probability Model

- Outcome measure in present demonstration: occurrence of a casualty
- Objective: develop model that identifies which concepts in complaints are associated with high likelihood of a casualty.
  - Example: In the context of the NHTSA complaint data, does brake failure increase the likelihood of a casualty?
- Flags developed using NHTSA data:
  - Accelerator Sudden
  - Air Bag
  - Another Vehicle
  - Brake Failure
  - Brake Pedal
  - Engine Failure
  - Fuel System
  - Manufacturer Defect
  - Safety Issue
  - Steering Failure
  - While Driving

Development of a Probability Model

- Steps:
  - Starting model: logit analysis limited to structured data
  - Identify flags that meet a minimum-frequency threshold
  - Evaluate minimum-frequency flags for use in the prediction model
  - Expanded model: minimum-threshold flags in a logit analysis
  - Quintile analysis

Development of a Probability Model

- Starting model: logit analysis is limited to “Police Report” field in structured data

- Outcome Measure: 0/1 for occurrence of a casualty in a NHTSA reported complaint

- Structured data: 0/1 for a police report

Flag	Log-Likelihood Ratio	Logit Coefficient	Chi-Square Stat
Intercept	2132.2	-3.77	1266.4
Police Report	1459.9	3.61	630.1

Development of a Probability Model

- Quintile analysis: logit coefficients are used to estimate probability of a casualty
- Raw data: casualty = 6% of sample cases
- Logit analysis limited to structured data:
  - Estimated probability of casualty > 40% = 46% of sample cases

Count Row Pct Col Pct	Less than 0.20	0.41 – 0.60
No Casualty	3960 94.5 97.8	231 5.5 54.1
With Casualty	91 31.7 2.3	196 68.3 45.9

Development of a Probability Model

- Expanded model: flags that meet a minimum threshold are included in logit analysis

- Starting Flags developed using NHTSA data:
  - Accelerator Sudden
  - Air Bag
  - Another Vehicle
  - Brake Failure
  - Brake Pedal
  - Engine Failure
  - Fuel System
  - Manufacturer Defect
  - Safety Issue
  - Steering Failure
  - While Driving

Development of a Probability Model

- Identify flags that meet a minimum-frequency threshold

Flag	Frequency in Complaints	Meets Min-Freq
Accelerator Sudden	0.070	√
Air Bag	0.119	√
Another Vehicle	0.031	√
Brake Failure	0.031	√
Brake Pedal	0.082	√
Engine Failure	0.012	
Fuel System	0.028	
Manufacturer Defect	0.006	
Safety Issue	0.116	√
Steering Failure	0.020	
While Driving	0.224	√

Development of a Probability Model

- Evaluate the min-freq flags for use in the prediction model

Flag	Frequency in Complaints	Correlation with 'Casualty' Outcome
Accelerator Sudden	0.070	-0.004
Air Bag	0.119	0.276
Another Vehicle	0.031	0.010
Brake Failure	0.031	0.012
Brake Pedal	0.082	0.001
Safety Issue	0.116	-0.069
While Driving	0.224	0.024
Police Report (from structured data)	0.095	0.524

Development of a Probability Model

- Expanded model: flags that meet a minimum threshold are included in logit analysis

- Structured data: 0/1 for a police report

- Unstructured data: 0/1 flags for—
  - Accelerator Sudden
  - Air Bag
  - Another Vehicle
  - Brake Failure
  - Brake Pedal
  - Safety Issue
  - While Driving

Development of a Probability Model

- Results from step-wise logit analysis

Inclusion Order	Flag	Log-Likelihood Ratio	Logit Coefficient	Chi-Square Stat
start	Intercept	2132.2	-3.83	1124.3
1	Police Report	1459.9	3.29	458.6
2	Air Bag	1402.3	1.24	57.0
3	Safety Issue	1388.1	-1.20	10.5
4	Accelerator Sudden	1383.4	-0.59	4.3
----	Another Vehicle	----	----	----
----	Brake Failure	----	----	----
----	Brake Pedal	----	----	----
----	While Driving	----	----	----

Development of a Probability Model

- Raw data: casualty = 6%
- Estimated probability of casualty > 60%: casualty = 69%
- Estimated probability of casualty > 40%: casualty = 67%

Count Row Pct Col Pct	Less than 0.20	0.21 – 0.40	0.41 – 0.60	0.61 - 0.80
No Casualty	3972 94.8 97.8	166 4.0 65.9	4 0.1 66.7	49 1.2 31.2
With Casualty	91 31.7 2.2	86 30.0 34.1	2 0.7 33.3	108 37.6 68.8

Comparing Results from Simple and Expanded Models

Count Row Pct Col Pct	Less than 0.20	0.41 – 0.60
No Casualty	3960 94.5 97.8	231 5.5 54.1
With Casualty	91 31.7 2.3	196 68.3 45.9

Count Row Pct Col Pct	Less than 0.20	0.21 – 0.40	0.41 – 0.60	0.61 - 0.80
No Casualty	3972 94.8 97.8	166 4.0 65.9	4 0.1 66.7	49 1.2 31.2
With Casualty	91 31.7 2.2	86 30.0 34.1	2 0.7 33.3	108 37.6 68.8

Comparing Results from Simple and Expanded Models

- Inclusion of unstructured data has produced a means to improve the segmentation of casualty/no-casualty claims.
- Raw data: casualty = 6% of sample cases
- Results from logit analyses
  - Complaints where estimated probability of casualty > 40%:
    - Analyses limited to structured data: 46% of sample cases
    - Analyses including unstructured data: 67% of sample cases

SUMMARY OF PRESENTATION

1) General Types of Data in Property-Casualty Claim Files

2) Examples of “Real World” Unstructured Data

3) Processing Unstructured Data

4) Incorporating Unstructured Data into Data Analytics

Strong caveat: Statistics in this presentation are for a very limited number of narrowly-defined cases from USDOL and NHTSA public-access databases. The cases and statistics are intended to demonstrate the principles of processing and analyzing unstructured data and not for drawing conclusions or inferences concerning the subject matter of the data.