Machine Learning in the Age of Unstructured Data

Blueocean Market Intelligence

Presented by,
Durjoy Patranabish, Blueocean Market Intelligence
Setting the context

Current state of unstructured data

Business Scenario

Machine learning in unstructured data analytics

Blueocean Market Intelligence : 360 Discovery
Driven by data explosion, access to affordable computing & business imperatives unstructured data is on the rise …

**Volume of data is exploding**
10x increase Zettabyte

10X

40

2013 2020

**Significant reduction** in cost of acquiring, storing, managing data

Data storage cost USD/GB

4 40

10X

40

2013 2020

Ever increasing data is a key resource that can be “analysed” to *generate insights* that can help organisations to take better and faster decisions

- 800 mn+ daily active users
- 200 mn+ emails are sent every minute
- 300 hours of video uploaded every minute
- 1 Exabyte data stored in cloud, growing rapidly
- 497 million connected devices added in 2014 alone

- 40% of business executives complain that they have too much unstructured data and unable to interpret them*
- 79% of businesses believe that using insights from unstructured data will boost revenue*
- 51% say lack of available talent is one of the biggest impediments to making better decisions with big data^*


© 2015 Blueocean Market Intelligence
Organizations are now embracing machine learning ...

- Large scale deployments of Machine Learning beneficial in terms of *improved speed and accuracy*
- Understands non-linearity in the data and generates a function mapping input to output (Supervised Learning)
- Recommended for *solving classification and regression problems*
- Ensures *better profiling* of customers to understand their needs
... but should remember the 3 important data considerations

- Machine learning is exciting and offers numerous benefits, however organization must make important considerations before embarking on the machine learning engagement
- Although machine learning offers the capability to train and retrain the data from time to time, data considerations are important to ensure the model is effective and provides the right output
- Data to be fed into the training model need to be assessed on the relevance, assumptions, and time to ensure that the training model provides the optimum output

<table>
<thead>
<tr>
<th>Data Horizon</th>
</tr>
</thead>
<tbody>
<tr>
<td>- How often or quickly does the new data be part of the training model?</td>
</tr>
<tr>
<td>- Is it a time series problem or is an incremental learning problem?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Have the right assumptions regarding the relevance of data been made?</td>
</tr>
<tr>
<td>- Are older instances more relevant than newer instances?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Obsolescence</th>
</tr>
</thead>
<tbody>
<tr>
<td>- How long does it take before data becomes irrelevant to the model?</td>
</tr>
<tr>
<td>- Is the time for obsolescence dependant on the current data?</td>
</tr>
</tbody>
</table>
Business Scenario

Industrializing unstructured data analytics for Listening Intelligence
Is “the data” the problem?
What will it be: **manual** or **machine driven process**?
Data, data everywhere: cluttered, un-structured, un-organized
Input: Unstructured text data

Data Classification
Placement of topics and data into pre-defined taxonomies

Data Clustering
Topics are grouped based on meta data relationship
Extraction of customer experience insights from customer conversations using ML and NLP

**Business Challenge**

Need: **Automate, lower cost, increase efficiency**
- Consistent trending
- Relevant, timely, actionable Insights
- Industrialize the classification operation throughout the lifecycle

**The Data Challenge**

Need: **Accuracy, scale, confidence**
- Deep experiences=deep classification
- Overlapping experiences=overlapping categories
- Voice ambiguity=imperfect training
- Junk values=clean up
- Source variation (Internet, Email, Chat, Agent Notes)
Traditional Approach that we used initially failed in large scale implementation ...

Traditional/Manual Approach

• Standard statistical tools were used to pre-process the data
• A comprehensive Natural Language Processing based rule set was designed
• The developed set was then fed into the classification engine
• QA testing was performed
Auto-code engine enabled scalability offering economies of scale ...

Failings of the Traditional/Manual Approach to text classification

- The traditional/manual approach to text classification falls short when building a scalable model
- Also in case of new scenarios and cases, the model might become outdated as it cannot adjust itself to accommodate the new data without manual intervention

Autocode Engine

- Python Engine is used for pre-processing of data
- NLP techniques like stemming and lemmatization were used to eliminate unimportant words
- Important words (features) are extracted based on their importance
- Classifier is build based on these features
- QA testing is performed
These business rules* are then run on validation sets and accuracy is determined

Business rules will be designed on the development set by understanding patterns

Data Access & Loading

Data pre-processing

Feature Extraction

Ridge’s Classifier

Model Validation

Training sets are loaded into a secure environment within Blueocean premises

Classify the data points into different confidence measures*

Training set is then loaded and partitioned into development set and validation set

**Business rules** are the rules based on which data points are classified to the respective categories

Based on confidence measure a data point (comment) in the training set can be classified to *High confidence, Low confidence, and Medium confidence*
Comprehensive Natural Language Processing techniques will be used to capture categories.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemmatization and stemming</td>
<td>Identify related words by reducing to root word and grouping</td>
</tr>
<tr>
<td>Decomounding</td>
<td>Dividing compound words into simpler one’s</td>
</tr>
<tr>
<td>Phrasing</td>
<td>Convert common phrases or idioms to a single term</td>
</tr>
<tr>
<td>Tokenization</td>
<td>Divide block of texts (paragraph/email) to individual units</td>
</tr>
<tr>
<td>Synonym expansion</td>
<td>Expanding synonyms to improve categorization rate for user query</td>
</tr>
<tr>
<td>Entity extraction</td>
<td>Used for nouns and noun phrases</td>
</tr>
<tr>
<td>Key phrase extraction</td>
<td>Select and extract phrases for classification of textual data</td>
</tr>
<tr>
<td>Entity co-occurrence and strength</td>
<td>Determine entity strength and relationships</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>Quantifying textual data into positive and negative sentiments</td>
</tr>
</tbody>
</table>
Each term is typically assigned a numerical score, usually, incorporating features of the document and the overall document collection.

Features are important words in a model based on which classification is achieved.

A model is build using numerous features that enhances or offers maximum accuracy.

Feature extraction is the part of the process where the best features are extracted from the data set.

The greater the number of features the greater the number of categories that can be captured.

Importance is the weight of a certain combination i.e. how many times a combination has appeared.

Best features include the most important terms which makes the feature more competent (more accurate).

Feature extraction helps us in identifying important words in a text document.
Ridge Classifier is an advanced regression technique which is ideal for text classification exercises and classifies the texts to based on identified features and

- Ridge’s classifier provides formulas based on which classification happens in the training data.
- Ridge’s classifier extracts the probability that a given input belongs to each class in a list of classes and upon identifying features classifies the text to identified categories.

Reason we use Ridge’s Classifier:

- Training time is less when compared to other algorithms.
- It accommodates most of the variables to be considered for the classification problem compared to other methods.
- Testing the model is more faster when we use Ridge’s classifier.
• Machine learning is a gradual process where the machine has to be trained on how to identify new and upcoming data

• The more training a model receives the better it is able to predict and classify the upcoming text into respective categories

• In case any new category has to be introduced, the model has to be retrained with the inclusion of the additional training data

• Training of new data will go through the same steps as described previously (depicted above)

• Post training, model accuracy would be determined on datasets not used for training
Our auto-coding engine transformed the way insights were generated and the organization saw tangible impact in their business.

**Business Impact on Industrializing Text Classification Operations**

**Strategic and Quality Management Team**
- 20% improvement in the efficiency
- Configured to optimize whatever metric drives the business
- The model has the capability to infinitely scale up to handle increased data rates

**Customer Experience Team/ Customer Support**
- More timely data availability
- Reduced costs by a factor of 400 times per coded/classified case
- No manual intervention is required at any stage
- Can operate in real time with the capability to react in milliseconds

**Product Reengineering/Product Development**
- Aggregation of top issues
- Justify, and prioritize customer experience changes
  - Identify upcoming trends based on
Business impact

Industrialized Text Classification results:

- 20% efficiency improvement
- 400x Reduced costs /coded case
- 2% to 100% inclusion
- 4x data sources
- 40% to 70% accuracy improvement
- Days to hours turnaround
- Faster solution enablement
- Higher data confidence
The Intelligent approach

Machine learning and NLP for Business Listening Intelligence

Needs vs. Wants
Methodology
Structured Data Patterns
Learning and predicting
Evaluation & QA
ACTION

Approach
Classification
NLP techniques
Classifier
Data confidence
ROI

Connected and talking for ROI and customer impact

AUTOMATED UNSTRUCTURED TEXT ANALYSIS OPERATION

© 2015 Blueocean Market Intelligence
All problems are opportunities in disguise!!
The next wave in machine learning ...

<table>
<thead>
<tr>
<th><strong>Streaming Computational Model</strong></th>
<th>Machine learning from streaming data for real time predictive analytics e.g., Weather updates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incremental algorithms, and periodic retraining with a batch algorithm are methods of updating the model based on live stream data</td>
</tr>
<tr>
<td></td>
<td>Supervised training algorithm form the base for streaming computational model</td>
</tr>
<tr>
<td><strong>Ensemble Learning</strong></td>
<td>Ensemble learning algorithms effectively aggregate the outputs from a series of predictive analytics models to form a lone output.</td>
</tr>
<tr>
<td></td>
<td>Advantages of this approach is it combines different types of models and helps to coalesce their outcomes</td>
</tr>
<tr>
<td><strong>Bootstrap Aggregating (Bagging)</strong></td>
<td>Improves the precision of Machine Learning methods in regression models and other model types.</td>
</tr>
<tr>
<td></td>
<td>Works for unstable classifiers like neural networks, and decision trees</td>
</tr>
<tr>
<td></td>
<td>Approximate the target by approximating the weight of the function</td>
</tr>
<tr>
<td><strong>Deep Learning</strong></td>
<td>Increasingly popular variant of neural networks, with more than the typical two processing layers</td>
</tr>
<tr>
<td></td>
<td>Products that can understand and learn from the images, text, and video clogging the web</td>
</tr>
<tr>
<td></td>
<td>Associated with cognitive computing and is ideal for sets of big data.</td>
</tr>
<tr>
<td></td>
<td>Google’s acquisition of DeepMind a notable in the importance of deep learning in years to come</td>
</tr>
<tr>
<td><strong>Adversarial Training</strong></td>
<td>A research field that lies in the “intersection of machine learning and computer security”</td>
</tr>
<tr>
<td></td>
<td>Done with the aim of producing more secure learning algorithms</td>
</tr>
<tr>
<td></td>
<td>Mainly considers the algorithm to be poor, and then work to improve I to eliminate potential vulnerabilities during learning and classification</td>
</tr>
</tbody>
</table>
Machine learning completes the 360 discovery process..

- **Predictive Ability for customer lifecycle management**
  Helps our partners identify latest trends

- **Customer Experience Management**
  Time and cost for identification of customer and employee issues will be reduced and reduce attrition

- **Brand Monitoring**
  Help companies to keep a tab of the health of their companies brand image by analyzing trends over a period of time

- **Understand sentiments and current buzz**
  Identify the positive/negative impact of content publishing and online marketing strategy

- **Digital Research**
  Reduces time related to topics and document searches by grouping documents
Blueocean is part of the Cross-tab group of companies, each a leader in its space

**Leader in analytics and market intelligence solutions**
- 360 Approach: Deep insights through combination of multiple data sources
- Continuous investment in IP and innovations

**Leader in market research operations outsourcing**
- Strong process and technology orientation; combined with core functional expertise
- Data Collection across 65 countries and 30 languages

**Leader in smartphone metering to track device, app and website usage**
- Offers syndicated reports and custom analysis
- Build custom panels

**Leader in online panels in emerging markets**
- Focus on high quality and representative panels
- Ability to build and manage custom panels and communities

Blueocean is part of the Cross-tab group of companies, each a leader in its space.
True Intelligence comes from analytics and synthesis of all available data and information sources – a 360 approach

- Drive data driven decisions through institutionalization of 360 Discovery™
- Strong pool of internal IP providing value add to solve problems that are not effectively solved by existing technology
- Top notch talent supplemented with speed and scalability
- Proven experience with dozens of Fortune 500 corporations
- Global delivery model with 1000 analytics professionals across the US, EU & India
# Overview of our Analytics Services

<table>
<thead>
<tr>
<th>Services</th>
<th>Data Management</th>
<th>BI Reporting</th>
<th>Basic and Advanced Analytics</th>
<th>Unstructured Data Analytics</th>
</tr>
</thead>
</table>
|          | • Data Architecture, Model & Design  
          | • Data Security  
          | • Data Quality  
          | • Data Warehouse & BI  
          | • Master Data Management  
          | • Metadata Management | • Standard and Ad Hoc Reports  
          | • Interactive Executive Dashboards  
          | • Multi platform implementation (mobile, PC, Tablet etc.)  
          | • Reports with drill-down and drill-up capabilities... | Viable statistical/mathematical model for  
          |                                          | • Customer Segmentation  
          |                                          | • Trend Analysis  
          |                                          | • New Acquisition  
          |                                          | • Portfolio Management  
          |                                          | • Attrition/Churn Analysis  
          |                                          | • Loyalty Management  
          |                                          | • Market Mix Modeling  
          |                                          | • Product Analysis... | • Social Media Analytics  
          |                                          | • Customer Feedback Analytics  
          |                                          | • Sentiment Analysis  
          |                                          | • Query Classification  
          |                                          | • Fraud Detection  
          |                                          | • New product development... |
| Key IPs  | • Link platform (data consolidation)  
          | • TurfAI big data platform | • Smart device dashboarding | • Marketing Optimizer  
          |                                          |                                          | • Risk and pricing engine | • Auto-coding engine for unstructured data  
          |                                          |                                          |                                          | • TurfAI Machine learning platform |
| Technology Expertise | MySQL  
          | IBM  
          | Teradata  
          | Informatica  
          | Cognos  
          | Spotfire  
          | MicroStrategy  
          | SQL Server  
          | Tableau  
          | COBOL  
          | SAS  
          | WPS  
          | SPSS  
          | Python  
          | Anaconda |
Thank You

Durjoy Patranabish
Senior Vice President
Blueocean Market Intelligence

durjoy.p@blueoceanmi.com
Appendix

NLP Approaches Explained
An illustration of how NLP works – a step by step illustration

1. Stopword Removal

['Bug Loss signal', 'Glitch Call volume low', "Problem Product X won't turn on", 'Problem Freezing rebooting', 'Glitch Battery life erratic', "Problem Can't detect SIM", 'Bug Camera error', 'Bug LED notification working', 'Annoyance It makes rattling noise', 'Problem Poor audio quality', "Glitch Product X reboots charging, won't charge", 'Bug Problem parsing package', "Glitch Wi-Fi won't connect drops inexplicably"

- A stopword list is a set of words that should be excluded from the results of a tool.
- Typically stopword lists contain so-called function words that don’t carry as much meaning, such as determiners and prepositions (in, to, from, etc.).
Based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes

- Identify related words by reducing to root word and grouping

['Bug Loss signal', 'Glitch Call volum low', "Problem Product X won't turn on", 'Problem Freez reboot', 'Glitch Batteri life errat', 'Problem Can't detect SIM", 'Bug Camera error', 'Bug LEDnotif work', 'Annoy It make rattl nois', 'Problem Poor audio qualiti', "Glitch Product X reboot charging, won't charg", 'Bug Problem pars packag', "Glitch Wi-Fi won't connect drop inexplic""]
Lancaster Stemming

['bug loss sign', 'glitch cal volum low', "problem produc x won't turn on", 'problem freez reboot', 'glitch battery lif er', "problem can't detect sim", 'bug camer er', 'bug led not work', 'annoy it mak rattl nois', 'problem poor audio qual', "glitch produc x reboot charging, won't charg", 'bug problem pars pack', "glitch wi-f won't connect drop inexpI"]

- Reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form.
- Features a rule execution mechanism and externally stored rules
An illustration of how NLP works – a step by step illustration

4 Lemmatization

['Bug Loss signal', 'Glitch Call volume low', "Problem Product X won't turn on", 'Problem Freezing rebooting', 'Glitch Battery life erratic', "Problem Can't detect SIM", 'Bug Camera error', 'Bug LED notification working', 'Annoyance It make rattling noise', 'Problem Poor audio quality', "Glitch Product X reboots charging, won't charge", 'Bug Problem parsing package', "Glitch Wi-Fi won't connect drop inexplicably"]

- Identify related words by reducing to root word and grouping
- Obtains a single word that allows you to group together a bunch of inflected forms, while taking the context into account