Social Media Text and Predictive Analytics

Invited Talk at Big Data Everywhere

Cambridge, MA

June 23, 2015

Dr. Subrata Das
Machine Analytics, Inc., Cambridge, MA
www.machineanalytics.com
Outline

- Analytics and Unstructured Data
- Information Structuring
- Document Classification
- Applications and Case Studies
“… leverage data in a particular functional process (or application) to enable context-specific insight that is actionable.” – *Gartner*
Analytics Categorized

- **Descriptive analytics**
  - Current and historical look at organizational performance.

- **Predictive analytics**
  - Predicts future trends, behaviour and events for decision support.

- **Prescriptive analytics** (a.k.a. decision support)
  - Determines alternative courses of actions or decisions given the historical, current and projected situations, and a set of objectives, requirements, and constraints.

**Descriptive Analytics**: How have been the monthly sales for the past twelve months? Who are the most valuable customers?

**Predictive Analytics**: What are the projected sales for the next six months? Who are the customers likely to leave?

**Prescriptive Analytics**: What actions could be taken to increase the sales? What incentives can be offered to encourage customers to stay/prevent from leaving?

Company Offers Services

Sales Transaction Database
Analytics Landscape

Blue Sky
Research to be transitioned between 5-10 years (e.g. deep semantical analyses, universal prediction, decision recommendation)

Cutting-Edge
Incubate ongoing research for differentiation (e.g. graphical models, deep learning, generative modelling, probabilistic programming, active and semi-supervised learning, natural language processing, crowd and cloud analytics)

Low-Hanging Fruits
Leverage off-the-shelf well-proven AI/ML technologies (e.g. decision trees, neural networks, Bayesian networks, support vector machine)

Current Practice
Buy commercial-off-the-shelf statistical and business rule packages (e.g. SAS, SPSS, R, Data Miner, JBoss)
### What do we have to build on?

<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Overall Approach</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical</strong></td>
<td>Non-deterministic relationships between variables are captured in the form of</td>
<td>Test hypothesis, regression analyses, probability theory, sampling, inferencing, …</td>
</tr>
<tr>
<td></td>
<td>mathematical equations and probability distributions</td>
<td></td>
</tr>
<tr>
<td><strong>Artificial Intelligence (AI)</strong></td>
<td>Domain experts provide knowledge of system behavior, and knowledge engineers</td>
<td>Logic-based expert systems, fuzzy logic, Bayesian networks, …</td>
</tr>
<tr>
<td></td>
<td>develop computational models using an underlying ontology</td>
<td></td>
</tr>
<tr>
<td><strong>Temporal</strong></td>
<td>Linear/nonlinear equations specify behavior of stochastic processes or of dynamic</td>
<td>Autoregression, survival analysis, Kalman filters, Hidden Markov Models, Dynamic Bayesian Networks, …</td>
</tr>
<tr>
<td></td>
<td>systems as state transitions and observations</td>
<td></td>
</tr>
<tr>
<td><strong>Machine Learning (ML)</strong></td>
<td>System input/output behavior is observed, and machine learning techniques extract</td>
<td>Clustering, neural networks, and various linear, nonlinear, and symbolic approaches to learning, …</td>
</tr>
<tr>
<td></td>
<td>system behavior models</td>
<td></td>
</tr>
</tbody>
</table>
Model-based Analytics – a high-level view

Computational Models

Information Extraction

Structured Data

Unstructured Text

Inference Engine

Analytics

Situation Assessment
Unstructured Data Defined

- “Unstructured Data (or unstructured information) refers to information that either does not have a pre-defined data model and/or does not fit well into relational tables. Unstructured information is typically text-heavy but may contain data such as dates, numbers, and facts as well.” – Wikipedia, 2011

- “Unstructured data are machine- or human-generated information where the data do not easily conform to standard data structures (such as rows and columns with well defined schema) and where the understanding of the data is not readily accessible without human or machine based interpretation.” – Oracle, 2009

- “Unstructured data (or unstructured information) refers to masses of (usually) computerized information in which every bit of information does not have an assigned format and significance. Examples of “unstructured data” may include audio, video and unstructured text such as the body of an email or word processor document.” – Aberdeen Group, 2009
Big Unstructured Data

- **Volume**: Enormous volumes of data
- **Velocity**: Pace at which data flows in from sources like business processes, machines, networks and human interaction with social media, mobile devices, etc.
- **Variety**: Many sources and types of data both structured and **unstructured**
- **Veracity**: Biases, noise, and abnormality in data
What is Structuring?

Image

Text

Homer is sitting on a chair drinking beer.

An alternative:

Homer is drinking beer while sitting on a chair.

Representation

RDF: metadata and triplets
(homer, seating, chair)
(homer drinking, beer)

An alternative:

Relational table:

<table>
<thead>
<tr>
<th>Action</th>
<th>Person</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>seating</td>
<td>homer</td>
<td>chair</td>
</tr>
<tr>
<td>drinking</td>
<td>homer</td>
<td>beer</td>
</tr>
</tbody>
</table>

Increasingly suitable for machine processing

Audio and Video

Analysis of mood and arguments of the actors involved

Challenge is information extraction and representation

Increasingly suitable for human consumption

Increasingly suitable for machine processing

Challenge is information extraction and representation

More complex

Analysis of mood and arguments of the actors involved

Increasingly suitable for human consumption

Increasingly suitable for machine processing
aText: Analysis of Text

- Extracts information from text and represents in the form of RDF triples via NLP, named entity recognition, coreference resolution, stemming, and dependency relation extraction.
- Supervised and unsupervised categorization of text documents.
- Applications:
  - Sentiment Analysis
  - Social Network Analysis
  - Document Summarization
  - Semantic Search

Trial version is available
Outline

- Analytics and Unstructured Data
- Information Structuring
- Document Classification
- Applications and Case Studies
Natural Language Processing

- Tokenization
- Morphological analysis
- Part-of-Speech tagging
- Syntactic Parsing
- Semantic analysis

Structuring and abstraction from text surface
Tokenization

- Tokenization segments a sequence of character codes (the input text) into a sequence of basic tokens (mostly wordforms, but also punctuation symbols, numbers).

Example

- Input text: *He got the job in spite of not having all the necessary qualifications.*
- Tokenized text:
  
  \[ \text{He} \ | \ \text{got} \ | \ \text{the} \ | \ \text{job} \ | \ \text{in} \ \text{spite} \ \text{of} \ | \ \text{not} \ | \ \text{having} \ | \ \text{all} \ | \ \text{the} \ | \ \text{necessary} \ | \ \text{qualifications} \ | \ . \]

Techniques

- Regular expression, Finite state automata
Morphological Analysis

- Morphology is the study of the relation between word surface forms (i.e. words as they appear in texts) and their lexical forms (lemmas and morphosyntactic features).
  - Example: hands \(\leftrightarrow\) hand + NOUN_PL

- Morphological analysis assigns each token all its possible morphosyntactic readings (base forms, part-of-speech categories, morphological features such as number or gender, etc.).
## Morphological Analysis Example

<table>
<thead>
<tr>
<th>Sequence of tokens</th>
<th>Lemma (base form) + POS and morphosyntactic features</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>he + Pron + Pers + Nom + 3P + Sg</td>
</tr>
<tr>
<td>got</td>
<td>get + Verb + PastBoth + 123SP</td>
</tr>
<tr>
<td>the</td>
<td>the + Det + Def + SP</td>
</tr>
<tr>
<td>job</td>
<td>job + Noun + Sg, job + Verb + Pres + Non3sg</td>
</tr>
<tr>
<td>In spite of</td>
<td>in_spite_of + Prep</td>
</tr>
<tr>
<td>not</td>
<td>not + Adv + Neg</td>
</tr>
<tr>
<td>having</td>
<td>have + Verb + Prog, have + Aux + Prog</td>
</tr>
<tr>
<td>all</td>
<td>all + Det + Pl + Quant, all + Pron + NomObl + 3P + Pl</td>
</tr>
<tr>
<td>the</td>
<td>the + Det + Def + SP</td>
</tr>
<tr>
<td>necessary</td>
<td>necessary + Adj</td>
</tr>
<tr>
<td>qualifications</td>
<td>qualification + Noun + Pl</td>
</tr>
<tr>
<td>.</td>
<td>. + Punct + Sent</td>
</tr>
</tbody>
</table>
POS Tagging

- Selects one morphosyntactic reading for each token according to context (i.e. disambiguating the morpholosyntactic readings of words).

- Rule based method disambiguate grammatical constraints in order to eliminate incorrect tags depending on context
  - A verb cannot follow a determiner
  - An auxiliary must be followed by a verb

- Statistical and machine learning techniques
## Morphological Analysis Example

<table>
<thead>
<tr>
<th>Sequence of tokens</th>
<th>Lemma (base form) + POS &amp; morphosyntactic features</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>he + Pron + Pers + Nom + 3P + Sg</td>
</tr>
<tr>
<td>got</td>
<td>get + Verb + PastBoth + 123SP</td>
</tr>
<tr>
<td>the</td>
<td>the + Det + Def + SP</td>
</tr>
<tr>
<td>job</td>
<td>job + Noun + Sg, job + Verb + Pres + Non3sg</td>
</tr>
<tr>
<td>In spite of</td>
<td>in_spite_of + Prep</td>
</tr>
<tr>
<td>not</td>
<td>not + Adv + Neg</td>
</tr>
<tr>
<td>having</td>
<td>have + Verb + Prog, have + Aux + Prog</td>
</tr>
<tr>
<td>all</td>
<td>all + Det + Pl + Quant, all + Pron + NomObl + 3P + Pl</td>
</tr>
<tr>
<td>the</td>
<td>the + Det + Def + SP</td>
</tr>
<tr>
<td>necessary</td>
<td>necessary + Adj</td>
</tr>
<tr>
<td>qualifications</td>
<td>qualification + Noun + Pl</td>
</tr>
<tr>
<td>.</td>
<td>. + Punct + Sent</td>
</tr>
</tbody>
</table>
Syntactic Parsing

- Builds syntactic structures of sentences (words are grouped into syntactic constituents and/or interconnected with grammatical relations)
- Models of syntactic structures
  - Constituent-based structure (or phrase structure)
  - Dependency-based structure
Constituent-based Parsing

- Grouping words into hierarchical labeled constituents (phrases)
- Parsing result is a phrase structure tree where nodes are phrases/constituents and leaves are wordforms.
He got the job in spite of not having all the necessary qualifications.
Dependency-based Parsing

- Links words with grammatical relations called dependencies (subject, object, attribute, etc.)
- Parsing result is a dependency structure, a graph where nodes are words and arcs are grammatical relations between words
- Example:
  - He got the job in spite of not having all the necessary qualifications.
  - SUBJ(got, He), OBJ(got, job), DET(job, the), META(got, in spite of), PCOMP(in spite of, having), NEG(having, not), OBJ(having, qualifications), DET(qualifications, the), ATTR(qualifications, necessary), DET(the, all)
He got the job in spite of not having all the necessary qualifications.
Semantic Analysis

- Builds a formal representation of the meaning of sentences from their syntactic structures and the basis meanings of words

Semantic analysis:
- Named entity recognition
- Coreference resolution
- Semantic relation extraction
Named Entities

- Named Entities (NEs) usually refers to the set of person names, location names, organizations names etc. occurring in a text.
  - Dates and numerical data are often added to this set.
  - Named entities are also extended to basic semantic elements of a specific domain, e.g. gene names in biology.

- Example:
  
  Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as a United States Senator from Illinois, from January 2005 until he resigned after his election to the presidency in November 2008.

  A native of Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review…
Named Entity Recognition

- **Standard Ambiguities:**
  - Washington ←→ PERSON or CITY
  - Turner ←→ PERSON or CITY or ORG or Common noun

- **Metonymy (Semantic Shift):**
  - Location vs. Organization
    - The prestigious Wimbledon tournament takes place in **Britain**.
    - **Britain** presented a draft UN resolution last month.
Coreference Resolution

- Coreference is a relation between linguistic expressions that refer to the same entity.
- Occurs when the same thing is referred to with multiple and possibly different expressions in a document.

Pronominal cataphora: forward reference to President George W. Bush

Pronominal anaphora: backward reference to President George W. Bush

Nominal anaphora: Backward nominal references to President George W. Bush

By choosing Paul Wolfowitz for the post of World Bank president right after he nominated John Bolton US ambassador to the United Nations, President George W. Bush has signaled his determination to send his administration's hardliners to the forefront of the international arena.

... Despite his assurances, Wolfowitz does not come off as a specialist on poverty and international development issues.

... For his part, Bolton, who will defend the US administration's foreign policy at the United Nations, has at times in the past been tough on the world body. "There is no such thing as the United Nations," he stated in 1994.

... Bush said Wednesday that the United States and its European allies would seek UN Security Council action against Iran if Tehran rejected incentives to limit its nuclear programs. "The understanding is, we go to the Security Council if they reject the offer. And I hope they don't. I hope they realize the world is clear about making sure that they don't end up with a nuclear weapon," the US president said.
Iran hires scientist

Scientist is hired by Iran

Dependency Relations to RDF
Outline

- Analytics and Unstructured Data
- Information Structuring
- Document Classification
- Applications and Case Studies
Methods Categorized

- **Generative Methods**
  - Provides full probabilistic model of all variables (conditional and priors)
  - “Generative” since sampling can generate synthetic data points
  - **Examples**: Naïve Bayes, Hidden Markov Models (HMM), mixture Gaussian models, Bayesian networks, Latent Dirichlet Allocation (LDA)

- **Discriminative Methods**
  - Provides a model only for the target variable(s) conditional on the observed variables (directly estimate posterior probabilities or estimate function between input and output)
  - Only sampling of the target variables conditional on the observed variables
  - Focus computational resources on given task, thus better performance
  - **Examples**: linear and logistic regression, SVMs, NN, Nearest neighbor
Methods Categorized

- Supervised learning
  - If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised.
  - Ex: feed-forward backprop NN, decision trees, …

- Unsupervised learning
  - Instances are unlabeled
  - Ex: LDA, …

- Semi-supervised learning
  - Mixed and makes sense in reality
Stemming

- Stemming is the conflation of the morphological variants of the same word into a common stem.
  - “apply” is a common stem of “application”, “applied”, and “applying”.
- In most cases, the stemming (e.g. Porter algorithm) leads to an improvement of the classification performance.
Naive Bayesian Classifier

- **Naïve Bayesian Classifier (NBC)** is a type of Bayesian network to classify a set of unlabeled data instances into a set of classes.
- **Supervised bag-of-words approach**

Naive Bayesian Classifier

\[
\arg\max_{C_i} p(C_i | d) = \arg\max_{C_i} p(C_i) \prod_j p(v_j | C_i)
\]
Latent Dirichlet Allocation (LDA) is an unsupervised topic detection technique.

An LDA model is a fully generative graphical model in which every topic is a distribution over the terms of the vocabulary, and every document is a distribution over the topics. These distributions are sampled from Dirichlet distributions.

There are several methods developed for making inference in LDA, including variational inferencing, expectation maximization, and Gibbs sampling.
LDA Model

\[ \alpha \] \rightarrow \mathbf{\theta}_m \rightarrow \phi_k \rightarrow \mathbf{w}_{m,n} \rightarrow \mathbf{z}_{m,n} \rightarrow \mathbf{w}_{m,n}

\mathbf{\theta}_m \sim \text{Dir}(\alpha)

\phi_k \sim \text{Dir}(\beta)

\mathbf{z}_{m,n} \sim \text{Mult}(\mathbf{\theta}_m)

\mathbf{w}_{m,n} \sim \text{Mult}(\phi_{z_{m,n}})

Blei et al., 2002
Outline

- Analytics and Unstructured Data
- Information Structuring
- Document Classification
- Applications and Case Studies
TripAdvisor Collection

- TripAdvisor dataset built by Baccianella et al., consisting of 15,763 hotel reviews from the TripAdvisor Web site.
- 10,508 documents used for training and 5,255 used for test.
- Training set contains 36,670 unique words.

203223_3024338  Fantastic  For the money the location is fantastic. It is about 300 yards from a metro station. From the hotel you can walk to the colosseum. You can practically see it when you leave the hotel. Yes the rooms are small but the hotel has great character. I would definitely stay there again and would recommend it to others.  _PROS_Nothing _CONS_Nothing  5

274573_3994017  Perfectly okay  The Hotel Suisse was just okay. Daniela was very helpful but the other people at the front desk were not. We found their attitude less than desirable. The rooms were clean and spacious. Breakfast delivered to the room was a bit awkward but certainly doable. The location to the Spanish steps was quite helpful. Easy access to the Metro and a taxi stand. They also asked for a 1 night stay in cash which we thought was odd.  _PROS_Nothing _CONS_Nothing  3

253154_3638452  Terrible experience  The twin room booked turned out to be a single with single bed and camp bed. By our third night we had an ant infestation which the management were unwilling to deal with. Eventually I had to spray the room and wait till 1:30 am before being able to return. We left.  _PROS_Nothing _CONS_Nothing  1
Sentiment Analysis

Machine Analytics, Inc.
Dr. Subrata Das has founded Machine Analytics and is a member of the Machine Analytics services and training team. Subrata is a well-known expert in the fields of data fusion, computational intelligence, and business analytics. Subrata's technical expertise includes theory of probability of statistics, mathematical logic, probabilistic and other formalisms for handling uncertainty, symbolic argumentation, graphical modeling, particle filtering, natural language processing, intelligent agents, and a broad range of computational artificial intelligence and machine learning techniques.

Subrata spent two years in Grenoble, France, as the manager of over forty researchers in the document content laboratory at the Xerox European Research Centre. He guided applied analytics research and development in the areas of unstructured data analyses, machine translation, image processing, and decision-making under uncertainty. Subrata was one of the five-members in the high-profile Xerox task force Knowledge Work 2020, alongside colleagues from the Palo Alto Research Center (PARC), to explore a strategic vision of the future of work.

Before joining Xerox, Subrata held the Chief Scientist position at Charles River Analytics in Cambridge, MA, where he led many fusion and analytical projects funded by DARPA, NASA, and various branches within the US.

SUMMARY: Dr. Subrata Das has founded Machine Analytics and is a member of the Machine Analytics services and training team. Dr. Subrata Das is a well-known expert in the fields of data fusion, computational intelligence, and business analytics. Dr. Subrata Das spent two years in Grenoble, France, as the manager of over forty researchers in the document content laboratory at the Xerox European Research Centre.
In spite of a successful rocket launch on Saturday, Jan 10 at 4:47 a.m. ET, things didn’t quite work out as SpaceX had hoped, when it tried to make history with an experiment.

The Falcon 9 rocket lifted off as scheduled from Launch Complex 39A, carrying the Dragon cargo spacecraft on a routine mission to resupply the International Space Station.

What went up is doing fine. What came down didn’t: the cargo capsule, not the rocket. It splashed down in the Pacific ocean around 8 a.m. ET, 10 minutes after separation.

The company tried to land the first section of the rocket, which is 14 stories tall, back down gingerly and on its feet on earth, while the rest of the rocket continued on.

A floating landing pad was waiting out in the Pacific ocean for the booster. It made it there, but came down a little too hard, SpaceX founder Elon Musk said in a tweet.
Word Cloud

Machine Analytics, Inc.
Topic or Event Extraction
Social Media and Web Search

Web Search
- Search Engine: Bing
- Search Type: AND
- Search String:
- Recency: Any time
- Exclude:
- Start Index: 3
- Count: 25
- Batch Size: 5
- Proximity: 1

Input
- Enter keywords, number of pages and posts:
  - Ex: Mesothelioma:2:550
  - Mesothelioma:2:250
- OK | Cancel

Web Application
- Select an application: Amazon Review
- OK | Cancel

Social Media Searching
- Search Type: Choose...
- Search String: Enter search string...
- Screen Name: Choose...
- Start Date: 04.20.2015
- End Date: 04.20.2015
- Count: 1000
- Search | Cancel
Contact

sdas@machineanalytics.com

Thanks!
About Machine Analytics

- Provides analytics and data fusion consultancy services for clients in government and businesses.
- Develops customized solutions using a combination of in-house, commercial-off-the-shelf, and publicly available tools.
- Arranges on- and off-site analytics and data fusion training courses.
- Benefits include:
  - Increased business intelligence and enhanced decision support.
  - Greater process efficiency via automation, leading to significant cost savings and profits.