Using Text Analytics to Spot Fake News

Tom Reamy
Chief Knowledge Architect

KAPS Group

http://www.kapsgroup.com

Author: Deep Text

Text Analytics and Fake News

- "Civilization is in a race between education and catastrophe. Let us learn the truth and spread it as far and wide as our circumstances allow. For truth is the greatest weapon we have."
  - H. G. Wells
- OED word of the year for 2016: Post-Truth

Agenda

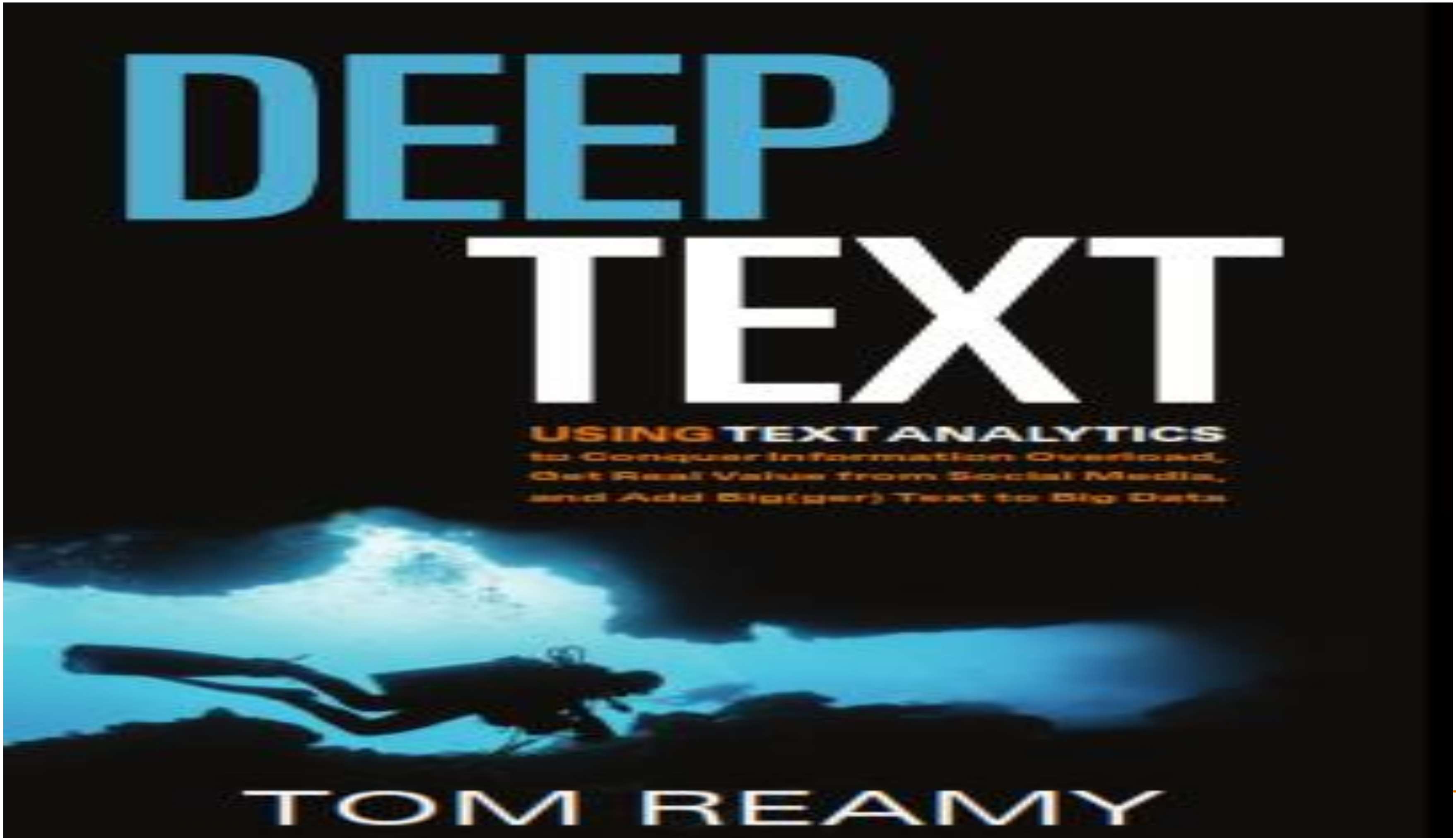
- Introduction
- Types of Fake News
- Techniques of Fake News
- Fake News in Context
- Proposed Solutions
- Solutions that Work
- Conclusion

Introduction: KAPS Group

- Network of Consultants and Partners – “Hiring”
- Strategy consulting – Text Analytics, Social Media, Integration
- Text Analytics Smart Start, Next Level
- Development - Taxonomy/Text Analytics, Social Media
- TA Train (1 day to 1 month)
  - Strategic personalized overview to hands on training
- TA Audit –Content, semantic resources, tech, info needs & behaviors
- Partners –Synaptica, SAS, IBM, Smart Logic, Expert Systems, Clarabridge, Lexalytics, BA Insight, BiText
- Clients: Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, Dept. of Transportation, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com

A treasure trove of technical detail, likely to become a definitive source on text analytics – *Kirkus Reviews* ***Information Today Table***

Book Signing at Reception – 17:15-18:30 Info Today table


DEEP
TEXT
USING TEXT ANALYTICS
to Conquer Information Overload,
Get Real Value from Social Media,
and Add Big(ger) Text to Big Data
TOM REAMY

Text Analytics and Fake News
What is Text Analytics?

- Text Mining – NLP, statistical, predictive, machine learning
- Entity / Fact Extraction
- Sentiment Analysis
- Auto-categorization
  - Training sets, Terms, Rules
  - Boolean– Full search syntax – AND, OR, NOT
  - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
- Deep Learning – neural nets – big and fast enough for patterns
  - Good on images, not concepts
- Deep Learning is a dead end – black box, tricks, fast data
  - No common sense and no strategy to get there

Text Analytics and Fake News
What is Fake News?

- Types of Fake News – sliding scale
  - Information out of context, Opinion, Misinformation
  - Alternative facts, Lies
- Fake people, automated bots
  - Twitter – most of top 20 accounts are bots – 1,300 a day
  - Hunter bots – impersonate people to discredit them
- Popularity – Google – can be manipulated
  - Search for Holocaust and get Neo-Nazi
- Two drivers: make money and manipulate people

Text Analytics and Fake News
What is Fake News? Stories

- Ad for Giuliani – support me because I took good care of my mistress
- Words change meaning – globalist for Jewish bankers
- Fine line between comedy/satire and fake news
- Las Vegas shooting – Google and Facebook – top stories were fake – some from known sites – liberal anti-Trump
- Search engines more trusted than regular news
- Algorithms designed to favor popular, get most likes and comments

Text Analytics and Fake News
Fake News Techniques

- “Tens of thousands of fraudulent Clinton votes found in Ohio warehouse”
  - Add: made up person – Randall Prince, electrician
  - Photo of a ballot box (from UK), Label person in photo as Prince
  - Add story details – plan to substitute boxes for real ones
- Got 6 mil views, generated $1,000 hr in ads
- Russians posted to protest a recent election – 2,000 bots attacked – complained posts were porno
- Twitter is worse – short, automated
- Confusion – fake news is news you don’t like

Text Analytics and Fake News
Fake News: In Context

- General – blurring lines opinion & facts, content and ads
- Growth of fringe groups – finding each other, finding “facts” that support them – from hate groups to science deniers
- Polarization and echo chambers
  - Confirmation bias
  - Network effects reward extremists, unfriend = bubble
  - Politics becomes like race/religion – won’t let my daughter marry a liberal
- The Internet is making us stupid
- This could be the first really major crisis of information age
- Getting worse – future = more info, more automation, virtual reality, blurring lines between real and imagined

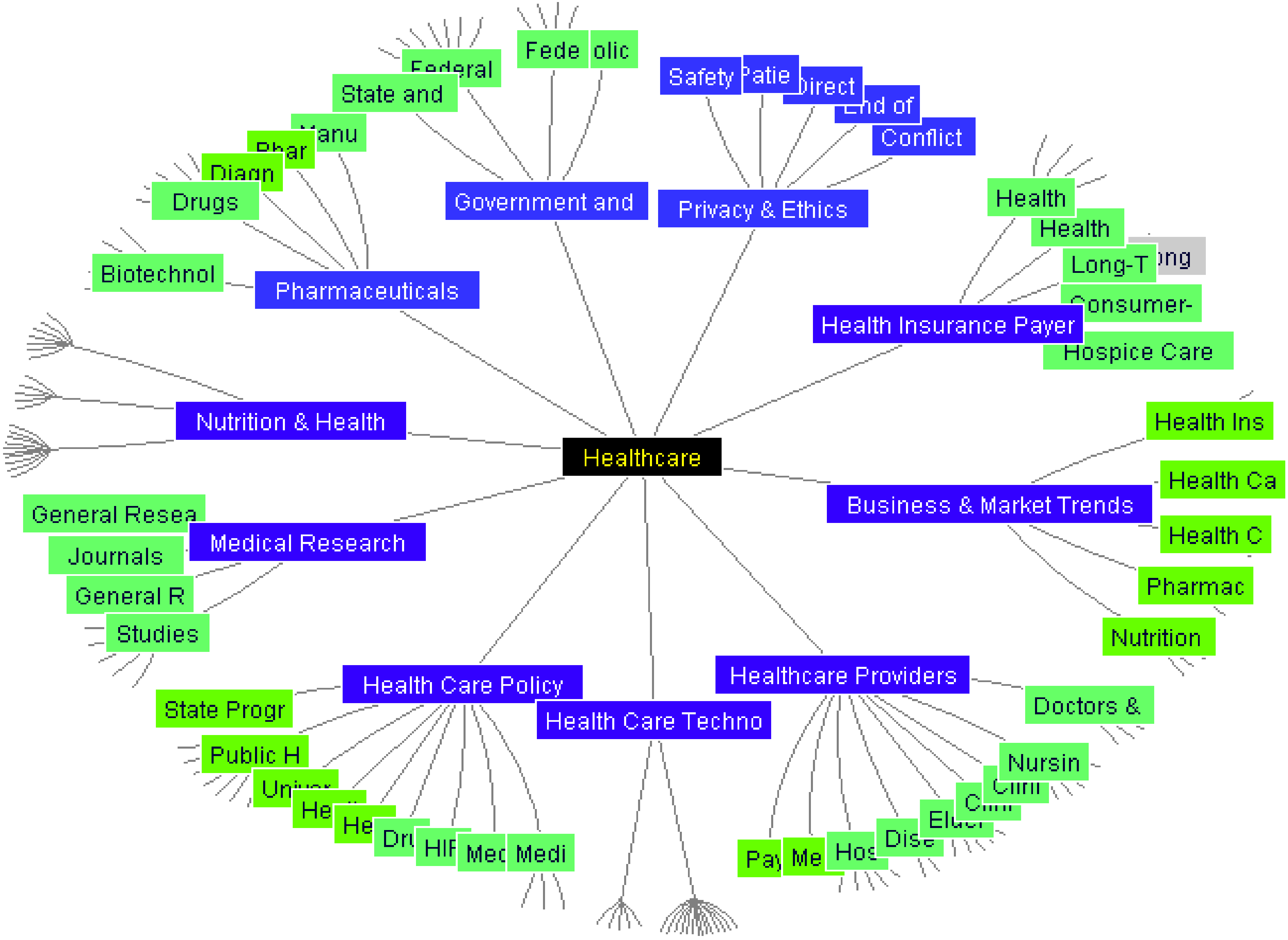
Text Analytics and Fake News
Proposed Solutions - Partial

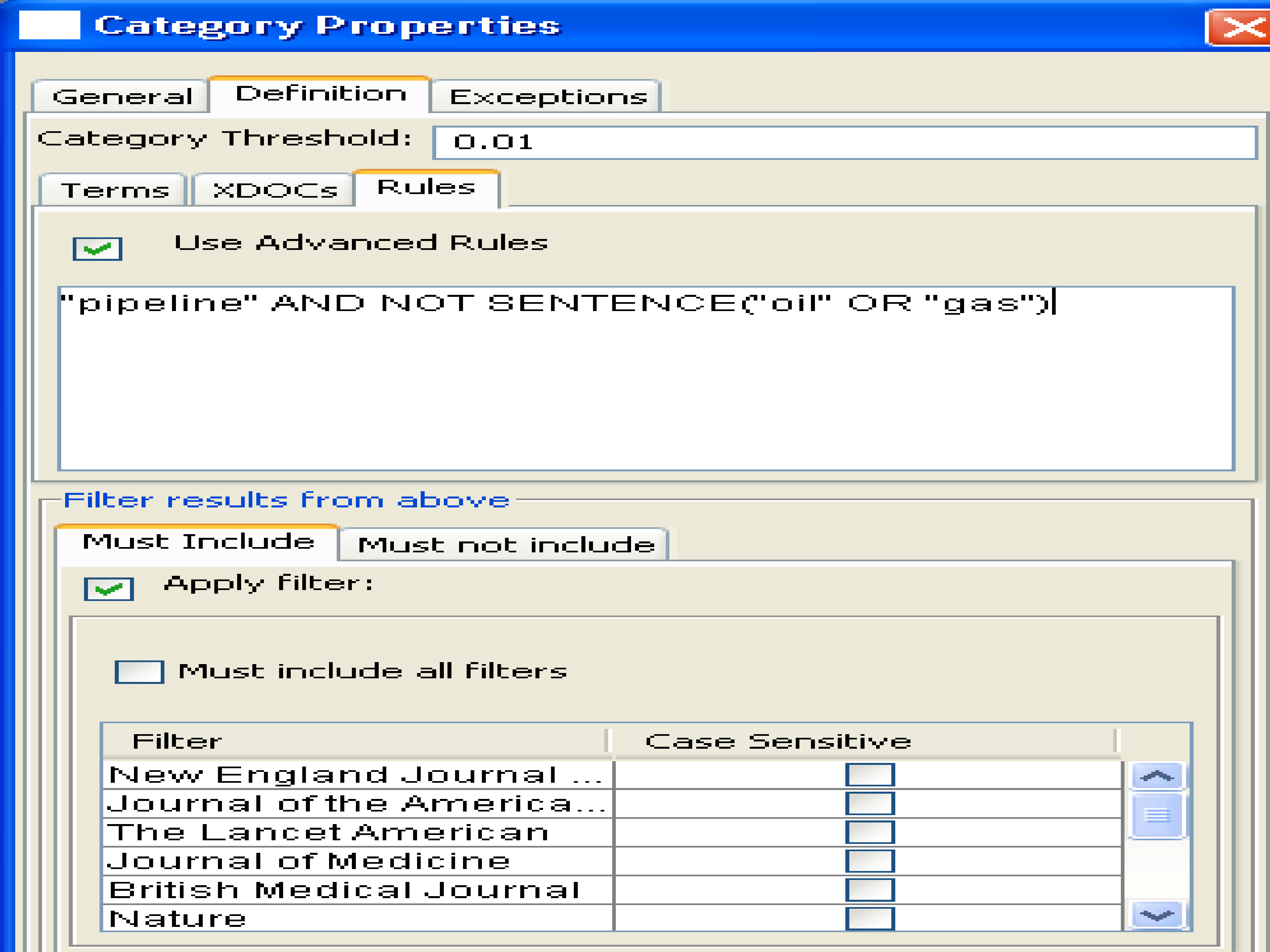
- Debunking
  - No money – fake news seen by millions, debunk = 1,000’s
  - Only facts, but arguments won with emotion and authority
  - Effects linger – George Lakoff – Don’t Think of an Elephant
- Financial: block ads
  - Advertisers not shunning like porno and gambling – yet?
  - Doesn’t deter political motivations
- Technical: tool to discover “sock puppets”, multiple sites/accounts

  Track and block known sites – URL based – abcnews.com.co, etc.
- Automated systems, machine learning, algorithms
  - Not smart enough (68% accuracy), can be manipulated
  - Black box – Watson – don’t know how it works

Text Analytics and Fake News
Case Study – Hybrid Analysis of News

- Inxight Smart Discovery
- Multiple Taxonomies
  - Healthcare – first target
  - Travel, Media, Education, Business, Consumer Goods,
- Content – 800+ Internet news sources
  - 5,000 stories a day
- Combination of Features – Categorization rules, Entity extraction, terms, Boolean, filters, facts
- Application – Editors get categorized results
  - Faster than human review
  - Smarter than automatic solutions


Healthcare
Pharmaceuticals
Biotechnol
Drugs
Diagn
Phar
Manu
State and
Federal
Government and
Fede
olic
Privacy & Ethics
Safety
Patie
Direct
End of
Conflict
Health Insurance Payer
Health
Health
Long-T
ong
Consumer-
Hospice Care
Business & Market Trends
Health Ins
Health Ca
Health C
Pharmac
Nutrition
Healthcare Providers
Doctors &
Nursin
Clini
Clini
Elder
Dise
Hos
Me
Pay
Health Care Techno
Health Care Policy
State Progr
Public H
Univer
He
He
Dru
HIP
Med
Medi
Medical Research
General Resea
Journals
General R
Studies
Nutrition & Health



Category Threshold: 0.01

☑ Use Advanced Rules

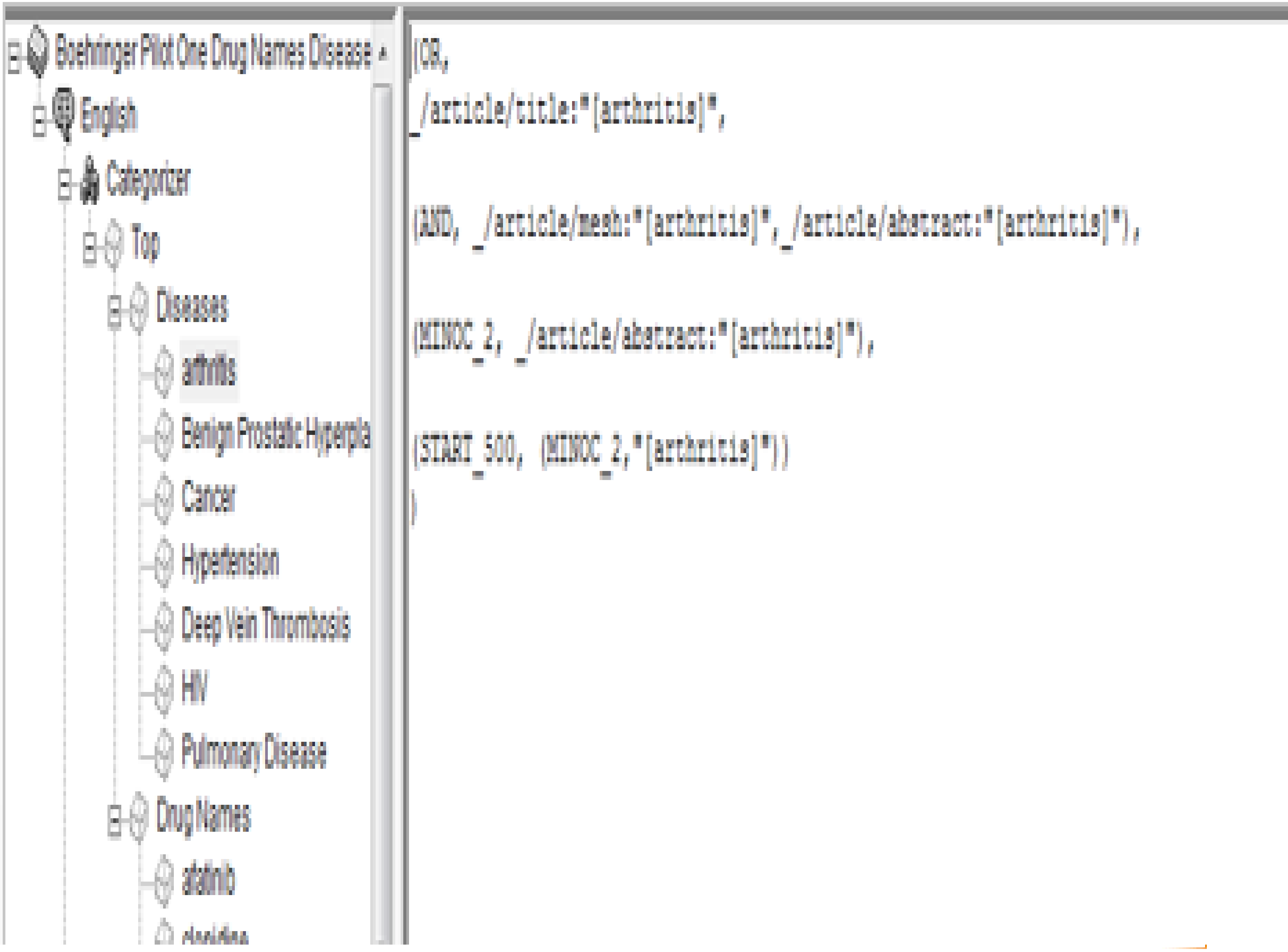
```
"pipeline" AND NOT SENTENCE("oil" OR "gas")
```

Filter results from above

☑ Apply filter:

□ Must include all filters

Filter	Case Sensitive
New England Journal ...	□
Journal of the America...	□
The Lancet American	□
Journal of Medicine	□
British Medical Journal	□
Nature	□



```
(OR,
_/article/title:"[arthritis]",


(AND, _/article/mesh:"[arthritis]",_/article/abstract:"[arthritis]"),


(MINOC_2, _/article/abstract:"[arthritis]"),


(START_500, (MINOC_2,"[arthritis]"))
)
```

Text Analytics and Fake News
Pronoun Analysis: Fraud Detection; Enron Emails

- Patterns of “Function” words reveal wide range of insights
- Function words = pronouns, articles, prepositions, conjunctions, etc.
  - Used at a high rate, short and hard to detect, very social, processed in the brain differently than content words
- Areas: sex, age, personality – individuals and groups, power-status,
- Lying / Fraud detection: Documents with lies have
  - Fewer and shorter words, fewer conjunctions
  - More use of “if, any, those, he, she, they, you”, less “I”
  - More positive emotion words
- Current research – 76% accuracy in some contexts
- Text Analytics can improve accuracy and utilize new sources

Text Analytics and Fake News
Deep Text Solution - All of the Above

- Need an all of the above approach – Technical, Financial, Linguistic, Categorical
- Mainstream news works very hard to validate – Facebook needs to do it too
- Facebook Initiatives
  - Need humans – adding 3,000 editors – hybrid solution
  - Also using external organizations – Politifact, Factcheck.org, Snopes
- Text Analytics – Meaning based
  - Depth of intelligence and speed of automated
  - Human-machine partnership = smarter humans.

Text Analytics and Fake News
Deep Text Solution – Filters and Fakeness Score

- Module 1 – database of known sites,
  - Block sites & ads
- Module 2 – Deep Learning – linguistic/social patterns
  - Function words, emotional intensity, abusive language
  - Writing style and posting activity
  - Poorer quality, shorter posts – often voted down
- Module 3 – Flexible categorization rules
  - Subject – political, controversial topics
  - Emotion and motivation taxonomies
- Fakeness Categorization Score – feed to humans

Text Analytics and Fake News
Solutions That Work

- All that helps but ultimate solution is education of society
- Who would believe Clinton had a child prostitute ring in a pizza place?
- Need to educate people to spot fake news and give them tools
- Real Time Debunking
  - Automated context
  - Linked Data – quick check – needs to be smarter
- Smart Reader / research assistant– see Deep Text
  - Automate some tasks, enrich others
- Ongoing war – as we develop better techniques, fakes will adapt

Text Analytics and Fake News
Resources

- Conferences – Information Today in Wash. DC
  - KMWorld, Taxonomy Boot Camp
  - Text Analytics Forum – New!
- Politifact.org / Google Fact Check / Factcheck.org
- Snopes – from urban legends to fake news
- Books:
  - Weaponized Lies: How to Think Critically in the Post-Truth Era by Daniel J. Levitin
  - Don’t Think of an Elephant by George Lakoff
  - Post-Truth: How Bullshit Conquered the World by James Ball

Text Analytics and Fake News
Conclusions

- Fake News is real and really serious
  - Undermine democracy, communication, civilization?
- Multiple factors driving more fake news – money, political, ease of technology & scale, it's "fun"
- Solutions require all of the above
  - Major initiatives from Facebook, Twitter, etc.
  - Multiple levels – technical, business, government
- Hybrid human-machine solutions – using text analytics
- Ultimate answer is better education

Questions?

: