AI for Unstructured Data in Finance

Machine Learning in Finance Workshop
May 17, 2019

Amanda Stent
NLP Architect, Office of the CTO

TechAtBloomberg.com
We Are in an “AI Revolution”

The Optimistic Promise of Artificial Intelligence
Andrew Ng and Tong Zhang on how AI is going to be like electricity, transforming every industry

Microsoft researchers achieve speech recognition milestone

China’s AI Agenda Advances

Google Translate Has Reached Human-Like Accuracy Thanks To Neural Machine Translation Engine

Alibaba's AI Outguns Humans in Reading Test

The Sublime and Scary Future of Cameras With A.I. Brains

Stanford team creates computer vision algorithm that can describe photos

TechAtBloomberg.com
© 2019 Bloomberg Finance L.P. All rights reserved.
The “AI Revolution” is Changing the World

Total number of mentions +% growth in number of mentions of "Artificial Intelligence" or "Machine Learning" in earnings call transcripts by quarter (1/1/07 - 12/31/17)
The “AI Revolution” is Data Driven

- Deep-learning based AI depends on:
  - Lots of specialized (tailored, expensive) compute
  - Lots of labeled data (supplied by you and me!)

Source: Michael Felsberg
Bloomberg is the Home of Market Data
Bloomberg Is the Home of Finance News

325K+ subscribers

Volume

2M stories per day
500 documents ingested per second

Index of 650M stories

News search

16M queries per day
Stories available for search in ~100ms
Average query response time <200ms

News alerting

1.5M subscriptions
500 stories matched per second
Alerts delivered in <100ms

TechAtBloomberg.com

© 2019 Bloomberg Finance L.P. All rights reserved.
News Moves Markets (Fast)

First Bloomberg Headline

New York Times Story

SEC Announcement

<20 minutes

12%
Social Media Moves Markets

Breaking: Two Explosions in the White House and Barack Obama is injured

Over $136 billion was wiped out in minutes

Over $136 billion was wiped out in minutes

TechAtBloomberg.com

© 2019 Bloomberg Finance L.P. All rights reserved.
Our Key Challenges

• The need for speed

• The need for precision

• The need for human in the loop
Text Analytics vs. NLP

Natural Language Processing (computational linguistics)
“the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language” (Stanford Encyclopedia of Philosophy)

Text Analytics
“the process of deriving information from text sources” (Gartner IT Glossary)
Why Do Text Analytics?

Most companies that use text analytics have one goal
- Assess customer feedback via voice, SMS, email… (e.g., to assess NPS)
- NLP tasks: entity recognition, sentiment analysis, key terms, visualization

A few have the goal of ranking/serving the text
- Treat text as a “first class object” (i.e., text is both input and output)
- NLP tasks: entity recognition and linking, topic detection, sentiment, clustering

A few have a different kind of goal
- Assess market signals (financial reports, news articles, tweets…) in order to predict some (financial) outcome (e.g., movie grosses, stock prices)
- NLP tasks: entity recognition and linking, sentiment analysis, relation/event extraction, signal extraction, classification
How Text Analytics Is Done

**Generation 0: write a bunch of rules (“templates”, “grammars”)**
- High-precision
- Slow, manual, difficult to maintain or update

**Generation 1: train a statistical classifier**
- For sequence tagging: conditional random fields; for document classification: linear/logistic regression, SVMs, decision trees/random forests
- Need labeled data

**Generation 2: deep learning and human in the loop**
- Need a lot of labeled data, or distant supervision
- May be slower

We still need separate “models” for each language/genre pair
- English news, English social, English generic; Chinese news, Chinese social…
The Life of a Document

- **Ingestion**
- **Content Extraction**
  - Text, tables, figures
- **Information Extraction**
  - Topics, entities, quantities, events, sentiment
- **Signal Extraction**
  - Counts of topics, entities, events… over time

**Key Insights, News Importance…**

**Ranking and Recommendation**

**Analytics**

**News Trends, Market Moving News…**
Topics are Central to AI for Finance

- Human curated in-house taxonomy of topic codes
- Exposed as an important metadata of text documents
Topics & Entities Convert Unstructured Content to Structured Data

- Over 1B tweets processed in the past 12 months
- Over 700 topic models deployed (still growing)
- Over 1M tweets labelled in-house for model training
- Peak-rate capacity of the real-time classification system

2016 U.S. Presidential Election
- ~100M tweets in 4 days
- peak rate at 4K+ tweets/sec
John Legere

I guess we’ll find out ;)

Walt Piecyk @WaltBTIG

Verizon’s post-paid phone subscriber base dropped 138k in Q1. T-Mobile might be the only one that grows this Q.

T-Mobile CEO Tweets on Phone Subscriber Base Speculation

By Christopher Kingston

(Bloomberg) -- BTIG analyst Walt Piecyk says on Twitter “Verizon’s post-paid phone subscriber base dropped 138k in Q1. T-Mobile might be the only one that grows this Q.”

- T-Mobile CEO John Legere embeds the analyst’s Tweet and says “I guess we’ll find out ;)”
- NOTE: Earlier, VZ said it lost 138k phone subs; reported overall new monthly subs of 565k, boosted by tablets
- NOTE: According to Bloomberg ests., TMUS gained 633.4k postpaid phone subs in qtr, AT&T gained 15.2k, Sprint lost 123k
Topics + Entities + Interaction = Insight

1. Manus Cranny: RT @UBS_CIO: Crude #oil: from bearish to sideways. Concerns about more potential US sanctions...
2. Mark Haefele: Crude #oil: from bearish to sideways. Concerns about more potential US sanctions on Iran and Ven...
3. MalaysianReserve: Oil market sees US sanctions on Iran a possibility
4. OilPrice.com: Barclays: Iran Could Find Other Buyers For Its Oil If Sanctions Tighten
5. Defend Democracy: The Petroyuan Is Here and Global Oil Markets Have Reacted Positively
6. NASDAQ: PDVSA plans April crude upgrader maintenance, more imports
7. WoodMac Oil Price: JCPOA exit, possible oil sanctions against Iran and war risk affecting oil prices today. Hit $7...
8. CNBC Wires: Oil prices rise on Middle East tension, global market recovery
9. Genepy Crude Oil: RT @baryallll: Any resumption of unilateral U.S. sanctions against Iran could drag down oil...
10. Asia Times (HK): China taking the long road to solve the petro-yuan puzzle
12. Baba Yaga: Any resumption of unilateral U.S. sanctions against the OPEC member could drag down oil exports fr...
13. Breaking News Forex: Crude Oil: WTI pulling back towards 64 as US sanctions imposed on South Sudan oil https:
14. Morningstar: Oil Rises on Falling Supply, Prospects of Iran Sanctions
15. Platts: Crude oil futures markets remain focused on geopolitical risk premia
(Topics + Entities + Market Fundamentals) / Time = Insight

<table>
<thead>
<tr>
<th>News Topic</th>
<th>A. Smith: German Beer Industry in Shock Over Monsanto Glyphosate Contamination (deadly mutagenic Roundup) ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Agricultural Chemicals</td>
<td>Livemint: RT @livemint_m2m: Is India turning the corner on usage of fertilizers? Urea sales till 20 March stood at...</td>
</tr>
<tr>
<td>2. Fertilizer</td>
<td>Skift: In parts of Iceland, you're more likely to meet a tourist than you are a local. The country doesn't want to ...</td>
</tr>
<tr>
<td>3. Iceland</td>
<td>Richard James: These Women Suffering Complications From Mesh Implants Were Let Down By The Health System, ...</td>
</tr>
<tr>
<td>4. Health Care Facilities &amp; Svcs</td>
<td>Manhattan Institute: @CPopeHC: What can the NHS learn from the American healthcare system, and vice versa? h...</td>
</tr>
<tr>
<td>5. Hospitals, Clinics</td>
<td>Hindustan Times: Astronomers are puzzled by a ‘see-through’ galaxy that doesn't have dark matter <a href="https://t.co/">https://t.co/</a>...</td>
</tr>
<tr>
<td>7. Env. Protection Agency</td>
<td>Govt of Pakistan: #AlbertEinstein #Quoteoftheday #QOTD #quote #Pakistan <a href="https://t.co/u3C72mVcq9">https://t.co/u3C72mVcq9</a></td>
</tr>
<tr>
<td>8. Pakistan Government</td>
<td>RT: Pakistani PM forced to undergo security check at US airport as Trump reportedly mulls sanctions against #ISL...</td>
</tr>
<tr>
<td>9. Pakistan Prime Minister</td>
<td>NYT National News: The Chilean government said it would start an inquiry into whether the remains of a tiny bab...</td>
</tr>
<tr>
<td>10. Chile Government</td>
<td>Bloomberg Markets: #5Things- North Korea and...Japan?- Tech drama- Indonesia’s new central bank chief- Upwa...</td>
</tr>
<tr>
<td>11. Korea Economy</td>
<td>ITWeb News Publication: Appeals court sends multibillion-dollar Java copyright case between Oracle and Google b...</td>
</tr>
<tr>
<td>12. Copyrights</td>
<td>Christopher Mahoney: &quot;Tesla's ratings reflect significant shortfall in the production rate of the Model 3. Tesla fac...</td>
</tr>
<tr>
<td>13. Hybrid Bonds</td>
<td>Calgary Herald: Egg Week: The absolutely official, completely objective, not biased at all, definitive ranking of w...</td>
</tr>
<tr>
<td>14. Canada</td>
<td>The African Voice: US President Donald Trump nominates Illinois Senator Kyle McCarter to serve as Ambassador o...</td>
</tr>
<tr>
<td>15. East African Community</td>
<td></td>
</tr>
</tbody>
</table>
(Extracted Signals + Market Fundamentals) / Time = Insight
(Extracted Signals + Market Fundamentals) / Time = Insight
Bloomberg in a Nutshell

Ingestion + Creation

Analyst Reports
Survey Data
Mails/Chats
Filings
News
Websites
Social Media
Ticker

Bloomberg

Analyses

Discovery
The Need for Precision

But without any other news that day to explain the move in the retailer’s stock price, some investors had a different theory: Computerized algorithms that trade stocks based on Twitter and social media alone had picked up on a surge in posts about Bill O’Reilly, and interpreted it as a signal to buy O’Reilly stock.
Human in the Loop

• In industry, no model is static
  o New entities
  o New contexts
  o New events and relationships

• Humans in the loop can:
  o Ameliorate lack of recall, precision
  o Provide important training data

• To enable hybrid workflow, more work is needed on:
  o Models “knowing what they know”
  o Active learning
  o Tools for humans to work with ML algorithms
Going Beyond Topics, Entities and Sentiment

Before

{Bearish, neutral, bullish}
Price target

Call

After

{Bearish, neutral, bullish}
Price target
Going Beyond Topics, Entities and Sentiment

Sundar Pichai, CEO: On the YouTube one. Look, I mean, the shift to video is a profound medium shift and especially in the context of mobile, you know and obviously users are following that. You’re seeing it in YouTube as well as elsewhere in mobile. And so, advertisers are being increasingly conscious. They’re being very, very responsive. So, we’re seeing great traction there and we’ll continue to see that. They are moving more off their traditional budgets to YouTube and that’s where we are getting traction. On mobile search, to me, increasingly we see we already announced that over 50% of our searches are on mobile. Mobile gives us very unique opportunities in terms of better understanding users and over time, as we use things like machine learning, I think we can make great strides. So, my long-term view on this is, it is as-compelling or in fact even better than desktop, but it will take us time to get there. We’re going to be focused till we get there.

Table 1: Earnings calls are extremely complex examples of naturally-occurring discourse. In this example question-answer pair from a Google earnings call on October 27, 2016, the analyst asks six distinct questions in a single turn. Because the interaction originates as speech, there are discourse markers and hedging. The analyst and executive discuss concrete entities and performance statistics and past, present and future performance.
Going Beyond Topics, Entities and Sentiment

- Bullish analysts are called on earlier in earnings calls
- Bullish analysts ask more positive questions; neutral/bearish analysts ask more negative questions
- Bullish analysts ask more concrete questions
- Neutral/bearish analysts ask about the past more
- Earnings calls are moderately predictive of changes in analysts’ forecasts
Some Recent Work


Data Science at Bloomberg

- **Our data:** We have the most text labeled with financial (*and legal!*!) entities, instruments and topics anywhere.

- **Our databases:** We have the most information about financial entities, instruments and topics anywhere.

- **Our communications:** We connect the most extraordinary individuals across global finance.

- **Our challenges:** Our business involves *every* AI challenge.
Where to Learn More

- TechAtBloomberg.com
  — Papers published in 2018 at EMNLP, SIGIR, NAACL HLT, WWW, and more…
  — Data for Good Exchange, Machines + Media, and other exciting events
  — Our work with JupyterHub, Solr, and other open source frameworks
  — Foundations of Machine Learning (FOML), David Rosenberg’s ML class

- GitHub.com/Bloomberg
  — bqplot: fast interactive plotting
  — sgtb: structured gradient tree boosting
  — foml: Foundations of Machine Learning
  — cnn-rnf: Convolutional Neural Networks with Recurrent Neural Filters
  — scatteract: extraction of labeled quantities from scatter plots
Thank You!

Questions?

TechAtBloomberg.com

© 2019 Bloomberg Finance L.P. All rights reserved.