DISSECTING SOCIAL MEDIA
AND ITS IMPACT ON FINANCIAL MARKETS

Machine Learning & Natural Language Processing at Bloomberg

James Hodson, R&D Knowledge Engineering
A BSV alert is triggered by an unusually high number of social media postings on a company.

1) INTEL CORP

<table>
<thead>
<tr>
<th>Actual</th>
<th>Expected</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2.2</td>
<td>Sep 13 09:00-09:30</td>
</tr>
</tbody>
</table>

### Number of Postings (30-Minute Increments)

- Alert
- Posts
- Expected
- INTC Equity 23.44

### Representative Postings

11) david peltier: Analysts' Actions: $AET $FINL $INTC $TEVA $UA thestreet.com/s... TWT 09/13
12) BenzingaWire: $INTC PreMarket Info Recap for September 13: Friday the 13th E... TWT 09/13
13) Earnings Impact: $INTC Check out the details about $300 Bay Trail clam shell... STK 09/13
14) Almostatrader: Watching: $NQ $SWY $INTC $GSVC... also names from earlier in... STK 09/13
15) John Voorheis: Upgrades $WEN $D $INTC $VZ $LUL $SWY $MGM Downgrades $UA... TWT 09/13
16) Charles Rankin: #stocks MARKET PULSE-Facebook, Intel, Pactera, Ulta Beauty,... TWT 09/13
We currently have a large position in APPLE. We believe the company to be extremely undervalued. Spoke to Tim Cook today. More to come.
Bloomberg

**Goal**
Use Data and Knowledge about the real world in order to infer patterns and make predictions. Leverage this ability to enhance the tools our clients need.

- Sentiment Analysis
- Novelty Detection
- Market Impact Analysis
- Topic-based Clustering
- Social Velocity
- Language Detection
- Event Detection & Extraction
- Machine Translation
- Social Graph Analysis
- Tokenization
- User Behaviour Analysis
- Parsing

James Hodson, R&D Knowledge Engineering
“James loves Machine Learning when the objective is well-specified.”
Mark the boundaries of tokens and sentences. Maximize the probability of each segment.

[“James loves Machine Learning when the objective is well-specified.”]
Identify the grammatical categorization of each token:

[ James loves Machine Learning when the objective is well-specified. ]
Noun groupings and named-entity recognition provides identification of logically connected entities. Named-entity resolution ties those entities to an ontology (knowledge-base) to differentiate them:

[ James loves Machine Learning when the objective is well-specified. ]

The ontology may also be used to resolve definitions of ambiguous words:

Verb, Present Indicative: to hold dear: CHERISH

James Hodson, R&D Knowledge Engineering
Build the dependency relationship diagram between constituents of the sentence:

[ James loves Machine Learning when the objective is well-specified. ]
What are the roles of different constituents in the sentence?

[Bloomberg]

[James loves Machine Learning when the objective is well-specified.]

James Hodson, R&D Knowledge Engineering
If we get this far, we can make our findings portable:

\[ \exists x \in X, \exists y \in Y : \text{hasObjective}(x,y) \land \text{wellSpecified}(y) \land \text{loves}(\text{James},x) \]

[ James loves Machine Learning when the objective is well-specified. ]

James Hodson, R&D Knowledge Engineering
What are the roles of different constituents in the sentence?

This is great if your content conforms to some core standards:

- Accepted grammatical constructs;
- Contextually consistent;
- Orthographic norms;
- Referentially consistent;
- Not “Kardashian Noise”;

[ James loves Machine Learning when the objective is well-specified. ]

James Hodson, R&D Knowledge Engineering
Bloomberg

Social Publishing
A new kind of content

Shared conversations: Stocktwits, Twitter, Facebook, Bit.ly, LinkedIn, more...

James Hodson, R&D Knowledge Engineering
Social Publishing
A new kind of content

James Hodson, R&D Knowledge Engineering
Bloomberg

Social Publishing
A new kind of content

Different Grammatical Constructs

Opinion Leaders
Different Grammatical Constructs
Inconsistent Entity References
Opinion Leaders

James Hodson, R&D Knowledge Engineering
Different Grammatical Constructs
Inconsistent Entity References
Contextually Ambiguous

Bloomberg

Social Publishing
A new kind of content

James Hodson, R&D Knowledge Engineering
Bloomberg

Opinion Leaders

Social Publishing
A new kind of content

Different Grammatical Constructs
Inconsistent Entity References
Contextually Ambiguous
Orthographically Challenging

James Hodson, R&D Knowledge Engineering
Different Grammatical Constructs
Inconsistent Entity References
Contextually Ambiguous
Orthographically Challenging
Frequency Spelling Inconsistencies

Social Publishing
A new kind of content

James Hodson, R&D Knowledge Engineering
In order to answer the same questions as we answer for editorial content, we must first address several new questions:

- **How trustworthy is this message?**
  - AP News Twitter account hacked: “Breaking: Two Explosions in the White House and Barack Obama is injured”
  - The false alarm sent the Dow plunging 145 points;

- **How authoritative is the source?**
  - Do people listen to what is being published?
  - Is it timely, novel, impactful?

- **What is the social context for this message?**
  - What has already been said?
  - Is this a response?
  - How many people are listening?
  - Is it spreading?
What does the content from this source usually look like?

- Normalize the input text to reduce dimensionality:

  - [Geeky, yet, brilliant], [!], [@modified@, @conversation@, A, card, game, based, on, lexicalized, tree-adjoining, grammar, @url@, @conversation@], [@topic@], [@topics@, grammar], [@topics@, SLPeeeps];

- Build a semi-lexicalized, feature-rich channel-based language model (LM);

- Estimate the likelihood that new content published is genuinely from this channel, by using a modified Query Likelihood Model:
  - \[ P(tweet|channel) = \sum_{k=0}^{n} \log(P(K|Q)) \]
  
  Where \( k \) represents the semi-lexicalized feature, \( K \) the set of n-gram features rooted at \( k \), and \( Q \) the semi-lexicalized channel LM.

James Hodson, R&D Knowledge Engineering
What does the content from this source usually look like?

- People tend to maintain a consistent set of styles across conversations.
  - Threshold of around 0.5 seems to separate genuine user content from spam behaviors;

---

MT: “ModifiedTweet”
@Frequent
#Consistency
N-gram Overlap in LM

James Hodson, R&D Knowledge Engineering
Is this genuine user content?

- Auto-generated template-spam from market activity;

James Hodson, R&D Knowledge Engineering
Are the assertions corroborated?

- The message syntax can provide us with clues;
  - Re-tweets imply agreement, propagation;
  - Citations from trusted sources (news);
- Have we seen this information before?
  - Simple n-gram overlap measures tend to perform poorly:

  ![Twitter screenshot 1]

  ![Twitter screenshot 2]

- Need to move beyond n-grams to entity recognition and propositional structure;

James Hodson, R&D Knowledge Engineering
Are the assertions corroborated?

- Dependency Parse of phrasal portions;
  - “UN Confirms” – X – “used”

- Super-sense mapping;
  - “sarın” >> WordNet.Noun.Substance
  - “chemical” >> WordNet.Noun.Substance

- FrameNet *Attack* frame;

James Hodson, R&D Knowledge Engineering
Do people listen to what is being published?

- Number of “followers”, “re-tweets”, alone don’t provide the full picture;
- Sources tend to be authoritative within a limited set of domains;
  - Don’t allow this to make everything they say important;

- We would like a topic-based score of a source’s authority:
  - First we need topics (LSI, LDA):
  - Or use #tags as proxy…
Do people listen to what is being published?

- Take Obama as an example;
- 6 months of posts give the following (roughly scaled by probability of each topic):

  - health-care
  - power
  - minimum-wage
  - energy
  - jobs
  - americans
  - discussion
  - business
  - insurance
  - pregnancy
  - marriage
  - debt
  - fuel-efficiency
  - shutdown
  - default
  - affordable
  - immigration
  - economy
  - environment
  - innovation
  - ohio
  - discrimination
  - against
  - mental-health
  - representatives
  - wind
  - bat-kid
  - steel
  - manufacturing
  - minimum-wage
  - carbon
  - pollution
  - innovation
  - fuel-efficiency
  - ohio
  - discrimination
  - representatives
  - wind
  - bat-kid
  - steel
  - manufacturing
  - minimum-wage
  - carbon
  - pollution
  - innovation
  - fuel-efficiency
  - ohio
  - discrimination
  - representatives
Do people listen to what is being published?

- Next, associate each source to their most prolific topics;
- \[ \arg\max_{\{t_1,...,t_n\}} [\prod_{d \in D} (\sum_{t \in \{t_1,...,t_n\}} P(d|t))] ; \]

Obama

James Hodson, R&D Knowledge Engineering
Do people listen to what is being published?

- For each topic, and each document representative of that topic for the source in question, we want to understand the relationship between prolific topics and their audience:
  - Build a document feature vector based on followers and re-tweets at the time of publication, trends of related #topics, URL’s, etc.
  - Use a PageRank-like algorithm:
    - Treat each user-topic as a node;
    - Each message is treated as an edge to the set of peers who interact with it directly or indirectly;
    - Relationship strength decays with distance and time;
Do people listen to what is being published?

- For Obama, the graph of topics vs. interaction profile—surprising result for healthcare, but potentially explained by temporal bias?
What is the social context for this message?

- Track conversations through the network;
- Analyze the strength of a topic by how quickly it emerges/grows/disappears;
- Identify when new important information is being released;
  - High authority score;
  - Unseen statement;
- Predict impact (information propagation through social graph);
- Predict sentiment (positive, negative, neutral)
- Verify impact (market reaction);

James Hodson, R&D Knowledge Engineering
Social Publishing
What is the impact?

James Hodson, R&D Knowledge Engineering
Questions:

jhodson2@bloomberg.net