Bloomberg

DISSECTING SOCIAL MEDIA

AND ITS IMPACT ON FINANCIAL MARKETS

Machine Learning & Natural Language Processing at Bloomberg

James Hodson, R&D Knowledge Engineering

GRAB

1<Go> for Company Information

Bloomberg Social Velocity - Sep 13 2013 09:21:02

A BSV alert is triggered by an unusually high number of social media postings on a company.

1) INTEL CORP   Actual 21   Expected 2.2   Range Sep 13 09:00-09:30

Number of Postings (30-Minute Increments)

10) Representative Postings

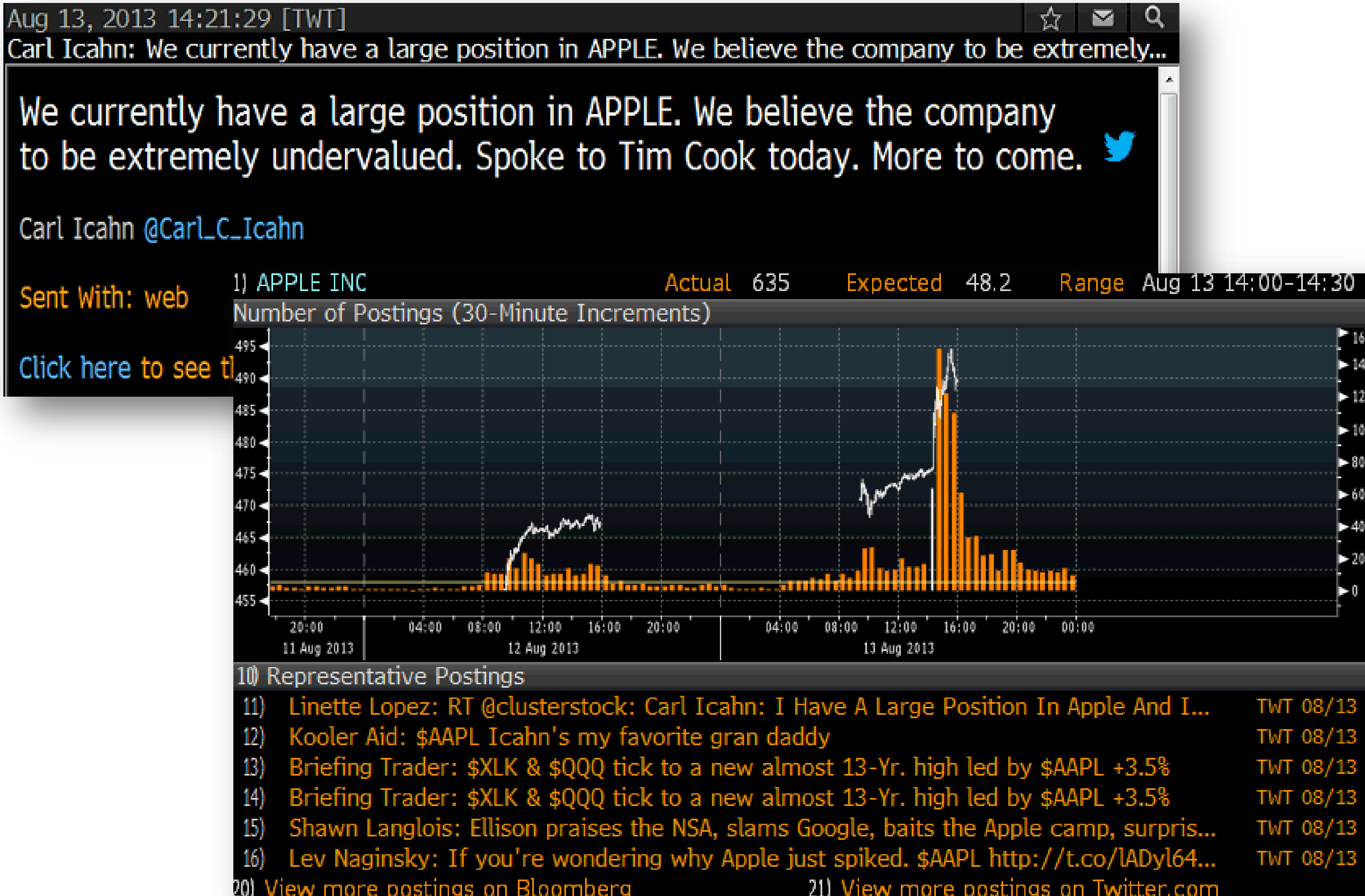
11)	david peltier: Analysts' Actions: $AET $FINL $INTC $TEVA $UA thestreet.com/s...	TWT 09/13
12)	BenzingaWire: $INTC PreMarket Info Recap for September 13: Friday the 13th E...	TWT 09/13
13)	Earnings Impact: $INTC Check out the details about $300 Bay Trail clam shell...	STK 09/13
14)	Almostatrader: 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

20) View more postings on Bloomberg   21) View more postings on Twitter.com

Bloomberg

Impactful

The world is watching



James Hodson, R&D Knowledge Engineering

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 Topic-based Clustering

Market Impact Analysis Social Velocity Language Detection

Event Detection & Extraction Machine Translation

Social Graph Analysis Tokenization

User Behaviour Analysis

Parsing

Conforming Content

Natural Language Processing

Start with some text.

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

Language

Encoding

Bloomberg

Natural Language Processing

Sentence and Token Boundaries

Mark the boundaries of tokens and sentences. Maximize the probability of each segment.

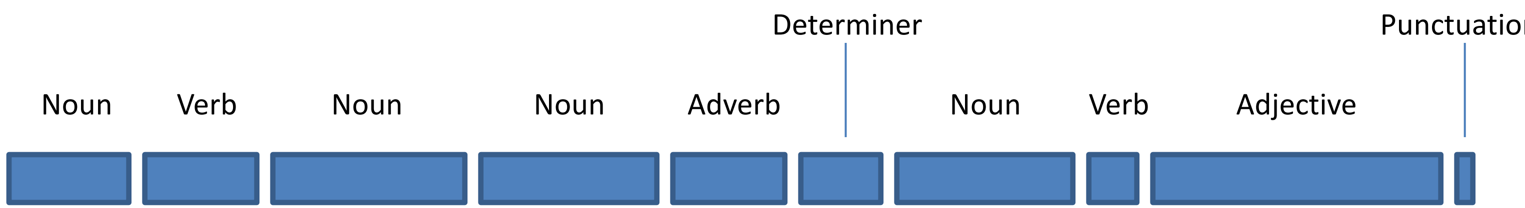
[“ James loves Machine Learning when the objective is well-specified . ”]

Bloomberg

Natural Language Processing

Parts-of-Speech Tagging

Identify the grammatical categorization of each token:



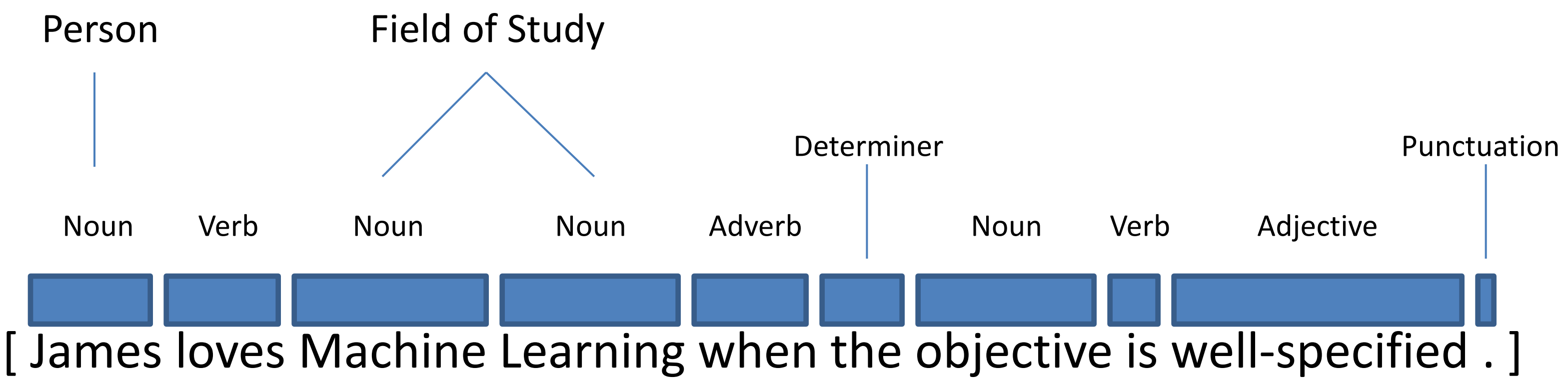
[James loves Machine Learning when the objective is well-specified .]

Bloomberg

Natural Language Processing

Named Entity Recognition

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:



Verb, Present Indicative: to hold dear : CHERISH

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

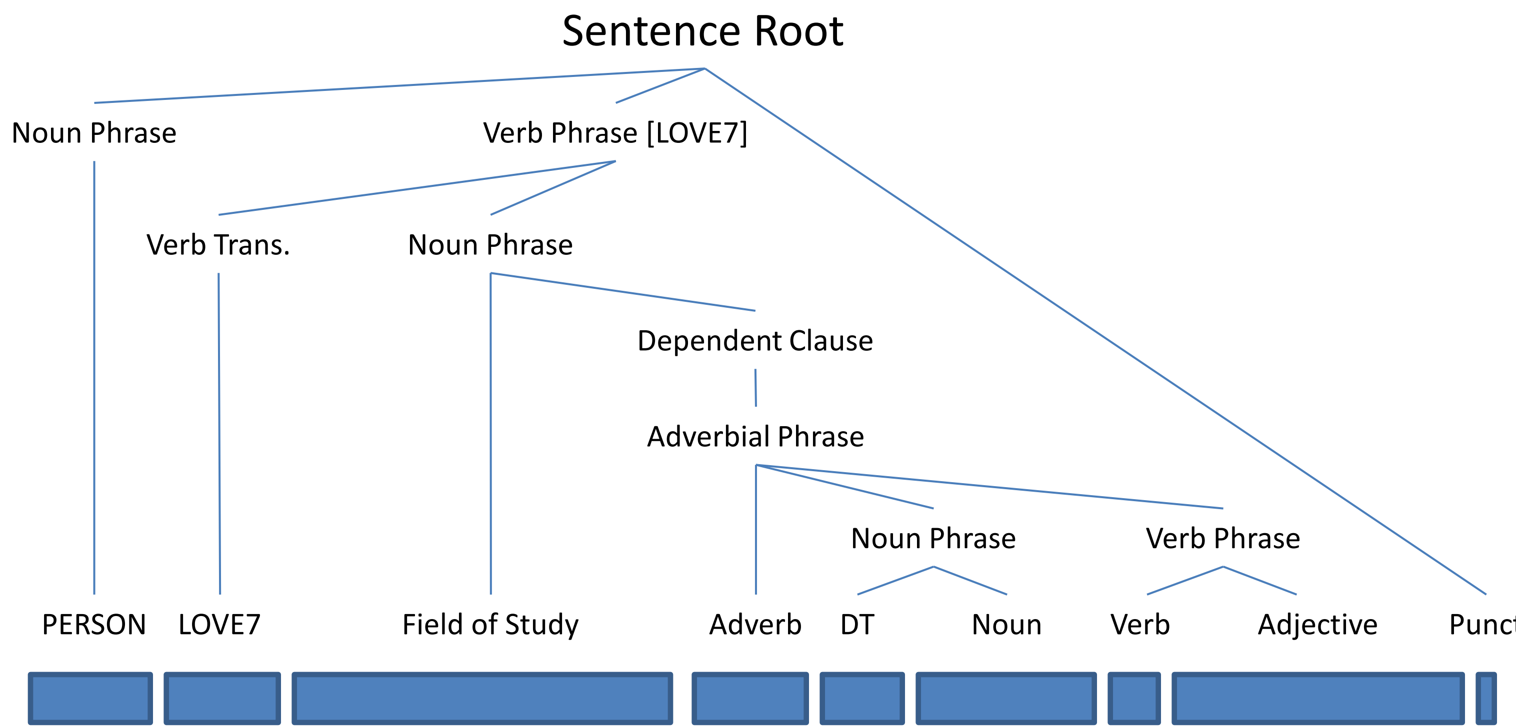
James Hodson, R&D Knowledge Engineering

Bloomberg

Natural Language Processing

Syntactic Parsing

Build the dependency relationship diagram between constituents of the sentence:



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

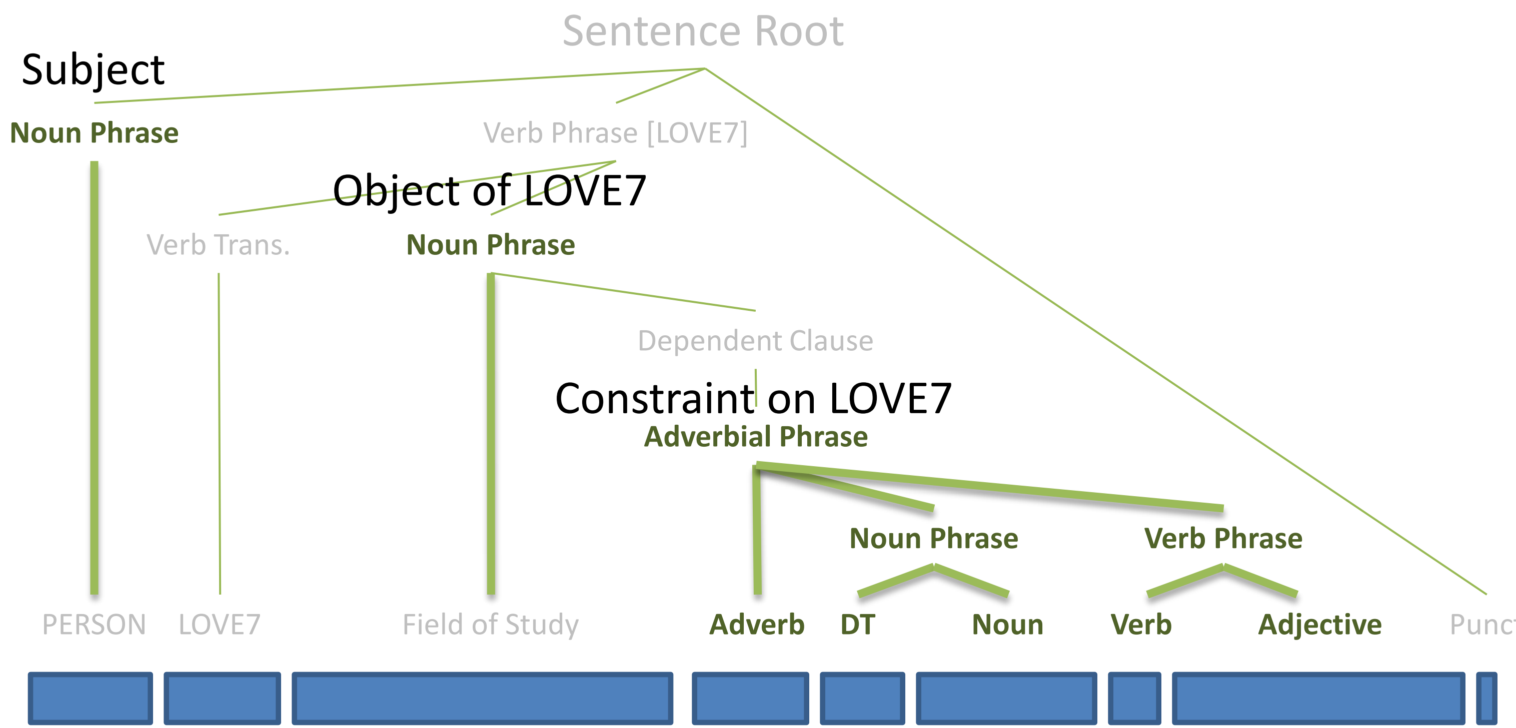
James Hodson, R&D Knowledge Engineering

Bloomberg

Natural Language Processing

Semantic Role Labeling

What are the roles of different constituents in the sentence?



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

James Hodson, R&D Knowledge Engineering

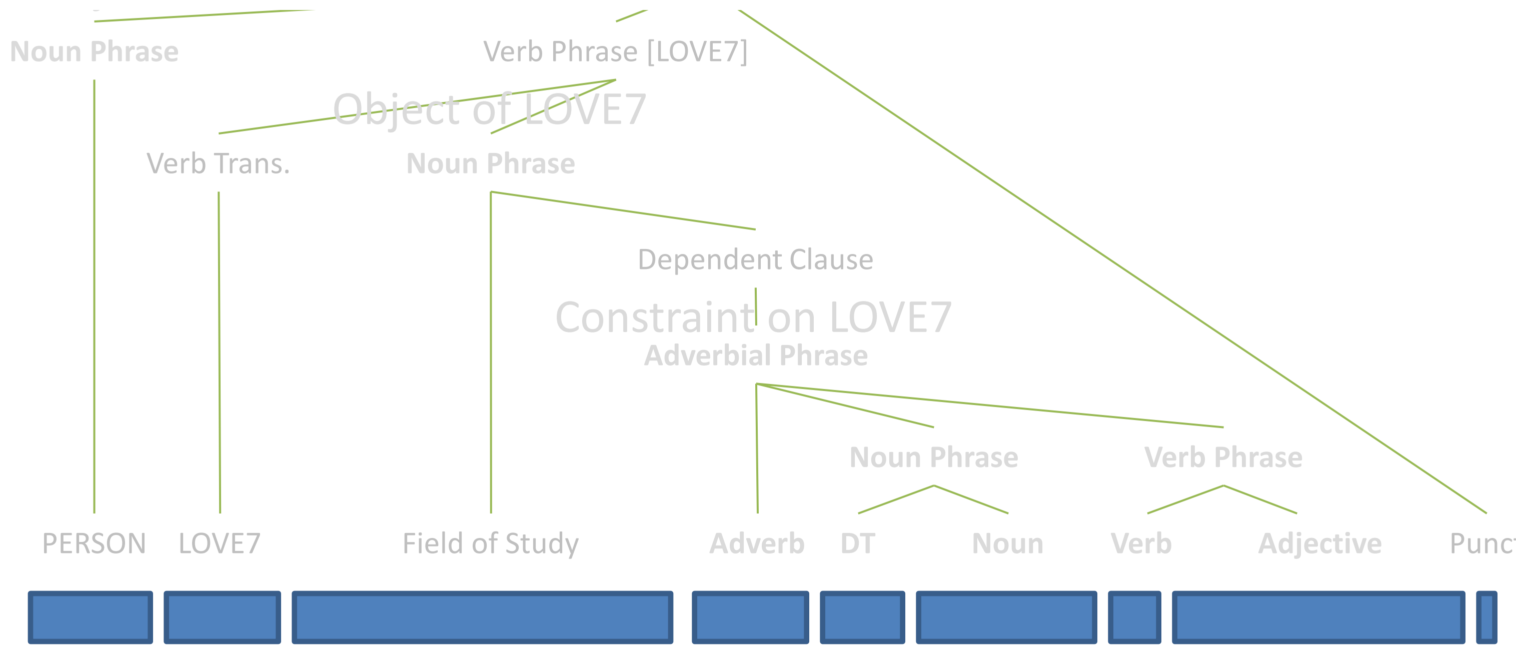
Bloomberg

Natural Language Processing

Representation

If we get this far, we can make our findings portable:

$$\exists x \in X, \exists y \in Y : hasObjective(x,y) \wedge wellSpecified(y) \wedge loves(James,x)$$



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

James Hodson, R&D Knowledge Engineering

Bloomberg

Natural Language Processing

Easy…?

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

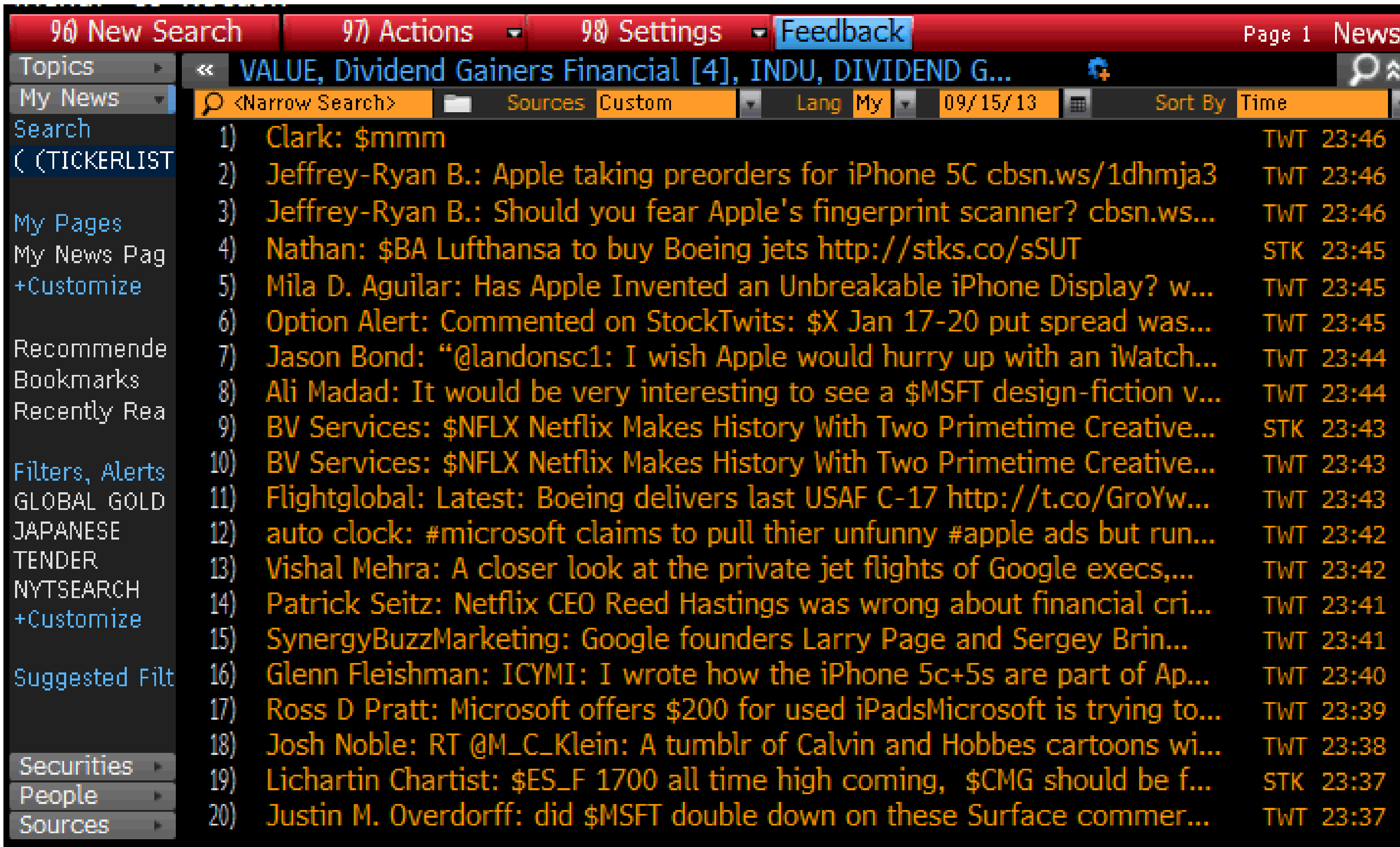
- Accepted grammatical constructs;
- Contextually consistent;
- Orthographic norms;
- Referentially consistent;
- Not “Kardashian Noise”;

Bloomberg

Social Publishing

A new kind of content

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



Bloomberg

Social Publishing

A new kind of content

Opinion Leaders

Bloomberg

Social Publishing

A new kind of content

Different Grammatical Constructs

Opinion Leaders

James Hodson, R&D Knowledge Engineering

Bloomberg

Social Publishing

A new kind of content

Different Grammatical Constructs

Inconsistent Entity References

Opinion Leaders

Bloomberg

Social Publishing

A new kind of content

Different Grammatical Constructs

Contextually Ambiguous

Inconsistent Entity References

Opinion Leaders

James Hodson, R&D Knowledge Engineering

Bloomberg

Social Publishing

A new kind of content

Different Grammatical Constructs

Contextually Ambiguous

Inconsistent Entity References

Orthographically Challenging

Opinion Leaders

Bloomberg

Social Publishing

A new kind of content

Different Grammatical Constructs

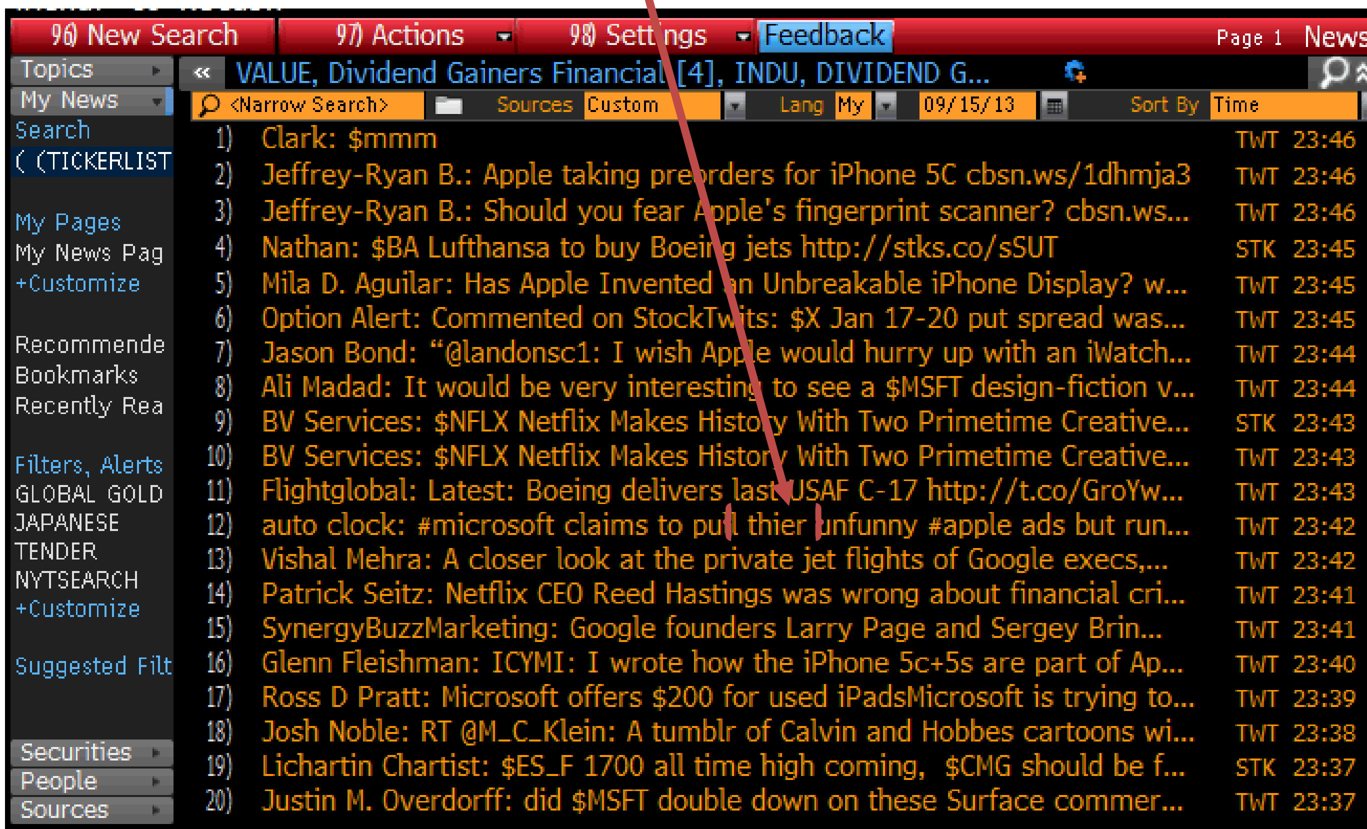
Contextually Ambiguous

Inconsistent Entity References

Orthographically Challenging

Opinion Leaders

Frequent Spelling Inconsistencies



Bloomberg

Social Publishing

A new kind of content

In order to answer the same questions as we answer for editorial content, we must first address several new questions:

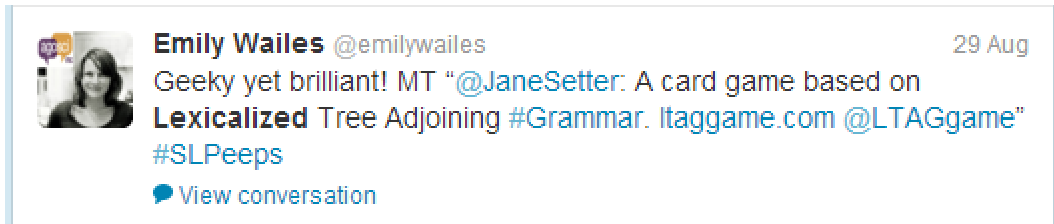
- How trustworthy is this message?
  - AP News Twitter account hacked: “Breaking: Two Explosions in the White House and Barack Obama is injured”
  - The false alarm sent the Dow plunging 145 points;
- How authoritative is the source?
  - Do people listen to what is being published?
  - Is it timely, 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?

Social Publishing

Is it trustworthy?

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@, SLPeeps];

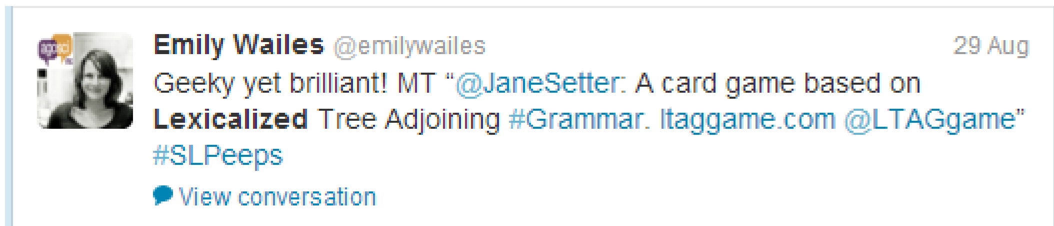
- Build a semi-lexicalized, feature-rich channel-based language model (LM);

- 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.

Social Publishing
Is it trustworthy?

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

Bloomberg

Social Publishing

Is it trustworthy?

Is this genuine user content?

- Auto-generated template-spam from market activity;

Bloomberg

Social Publishing

Is it trustworthy?

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:

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

Bloomberg

Social Publishing

Is it trustworthy?

Are the assertions corroborated?

- Dependency Parse of phrasal portions;
  - “UN Confirms” – X – “used”
- Super-sense mapping;
  - “sarin” >> WordNet.Noun.Substance
  - “chemical” >> WordNet.Noun.Substance
- FrameNet *Attack* frame;

Bloomberg

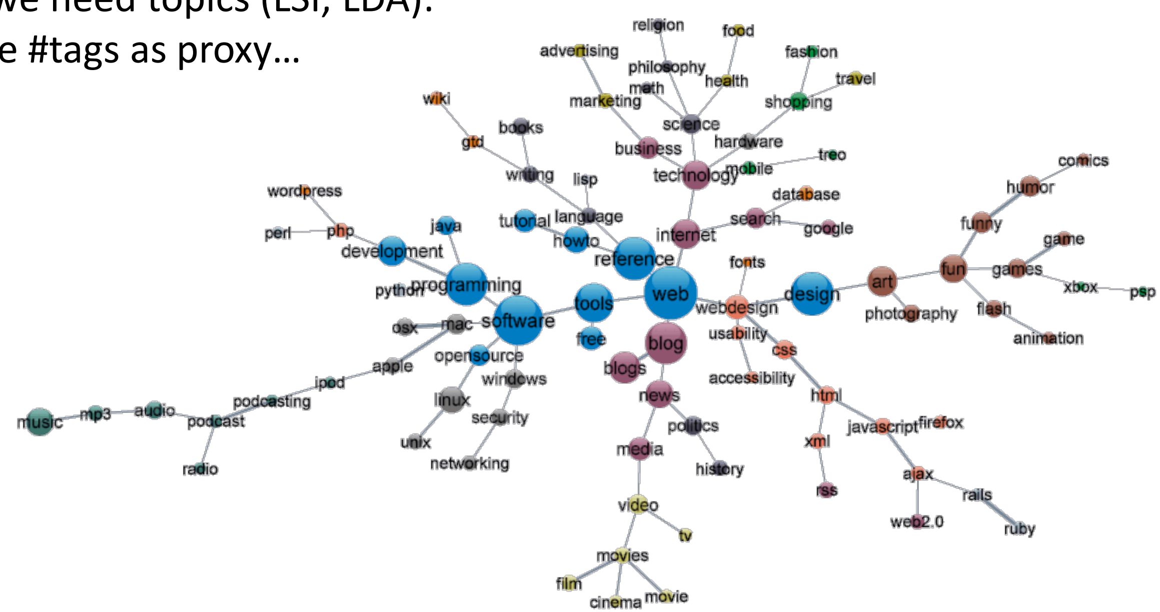
Social Publishing

How authoritative?

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…



Bloomberg

Social Publishing

How authoritative?

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 default
independence reform
jobs obamacare renewable manufacturing affordable
business steel
energy bat-kid
immigration
discussion carbon pollution
deficit
insurance discrimination
disaster against mental-health
pregnancy americans
representatives
ohio economy
marriage innovation
millions environment
patriots
debt shutdown
fuel-efficiency

James Hodson, R&D Knowledge Engineering

Bloomberg

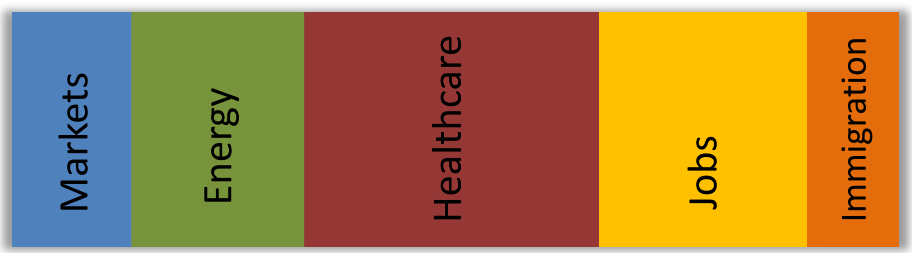
Social Publishing

How authoritative?

Do people listen to what is being published?

- Next, associate each source to their most prolific topics;
- $\underset{\{t_1,\ldots,tn\}\in T}{argmax_n}[\prod_{d\in D}(\sum_{t\in\{t_1,\ldots,tn\}} P(d|t))]$;

Obama



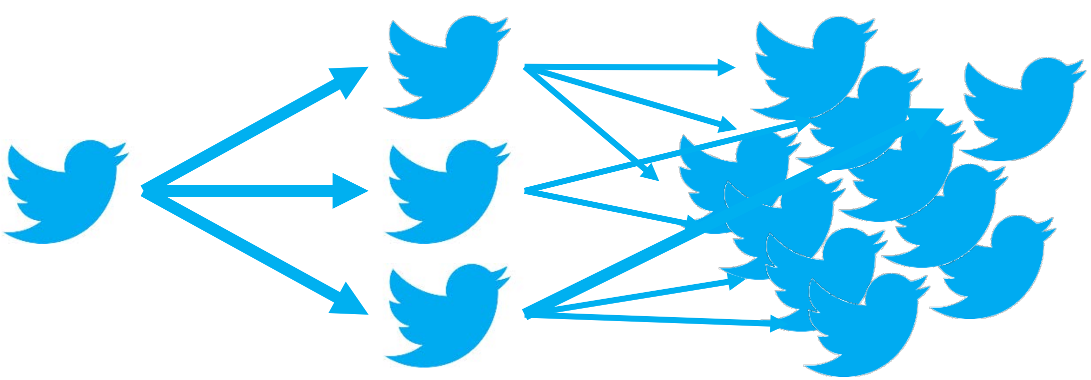
Bloomberg

Social Publishing

How authoritative?

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;



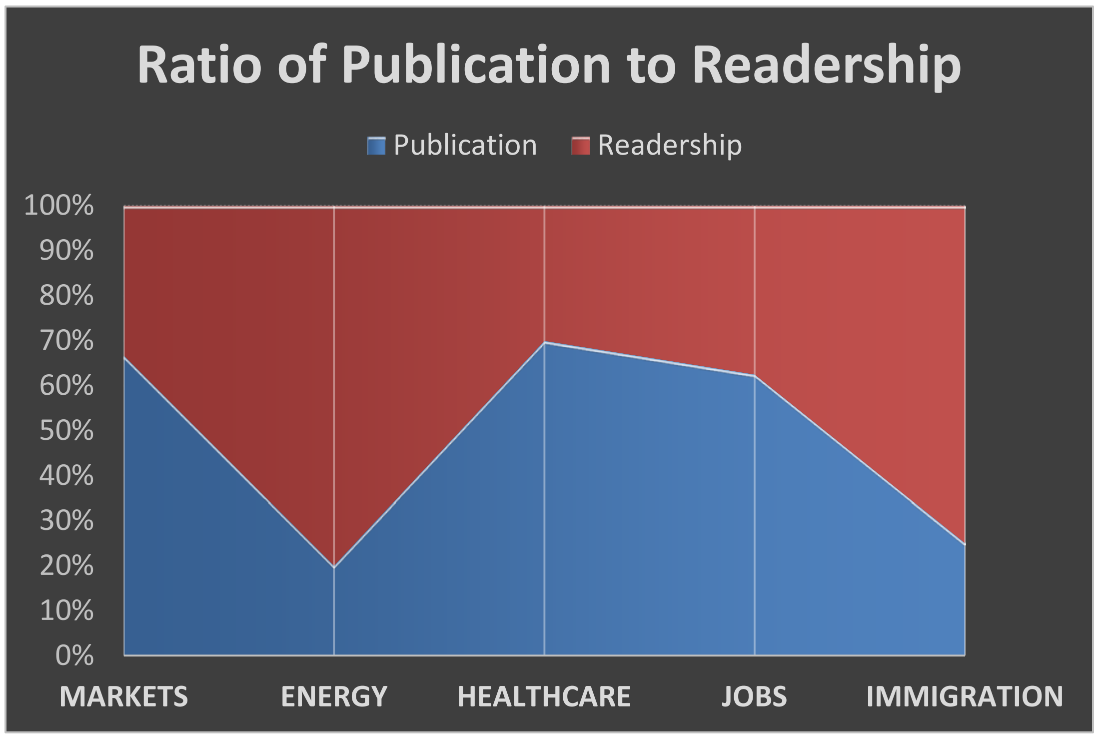
Bloomberg

Social Publishing

How authoritative?

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?



Social Publishing

What is the context?

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);

Bloomberg

Social Publishing
What is the impact?

Bloomberg

Questions:

?

jhodson2@bloomberg.net

James Hodson, R&D Knowledge Engineering