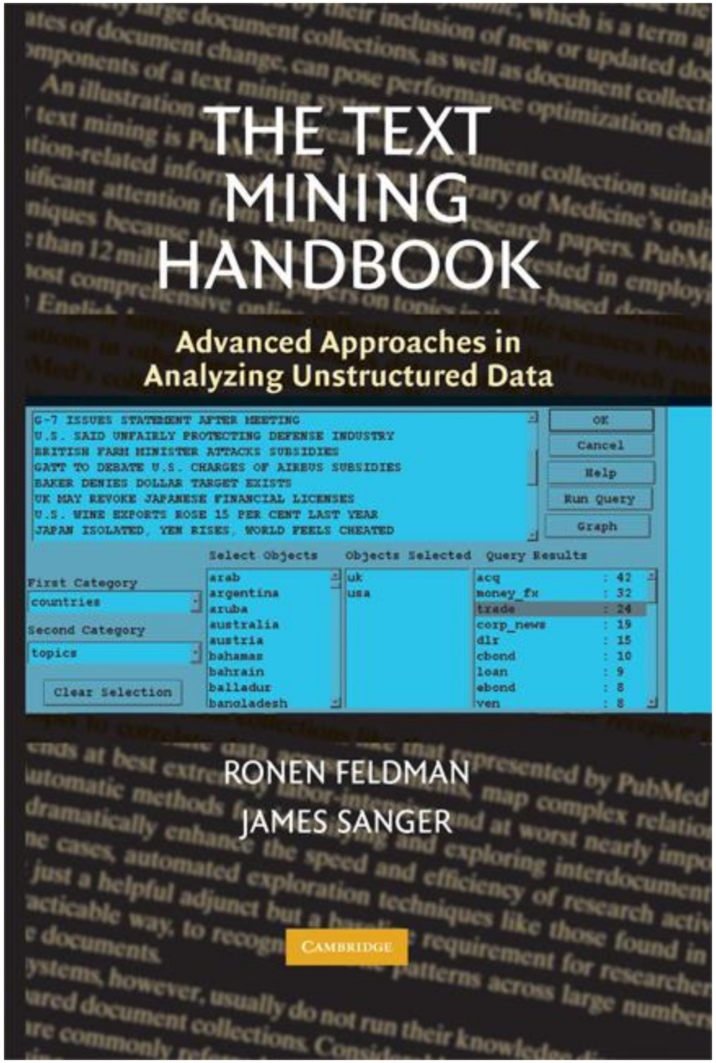
SENTIMENT ANALYSIS TUTORIAL

Prof. Ronen Feldman
Hebrew University, JERUSALEM
Digital Trowel, Empire State Building
Ronen.Feldman@huji.ac.il

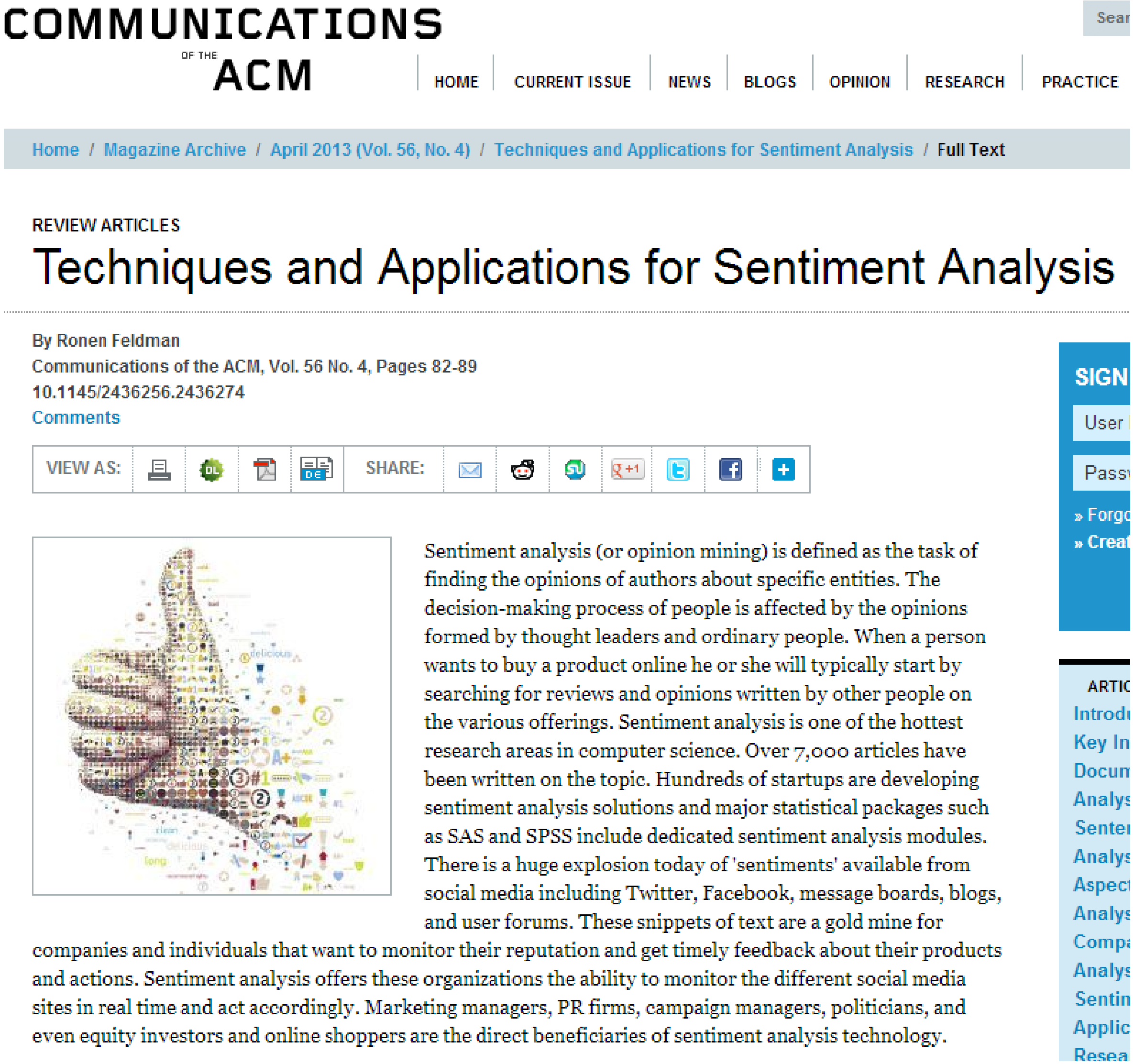
digital trowel
Pure Intelligence.

האוניברסיטה העברית בירושלים
The Hebrew University of Jerusalem

The Text Mining Handbook



CACM Article



REVIEW ARTICLES

Techniques and Applications for Sentiment Analysis

By Ronen Feldman
Communications of the ACM, Vol. 56 No. 4, Pages 82-89
10.1145/2436256.2436274
Comments

Sentiment analysis (or opinion mining) is defined as the task of finding the opinions of authors about specific entities. The decision-making process of people is affected by the opinions formed by thought leaders and ordinary people. When a person wants to buy a product online he or she will typically start by searching for reviews and opinions written by other people on the various offerings. Sentiment analysis is one of the hottest research areas in computer science. Over 7,000 articles have been written on the topic. Hundreds of startups are developing sentiment analysis solutions and major statistical packages such as SAS and SPSS include dedicated sentiment analysis modules. There is a huge explosion today of 'sentiments' available from social media including Twitter, Facebook, message boards, blogs, and user forums. These snippets of text are a gold mine for companies and individuals that want to monitor their reputation and get timely feedback about their products and actions. Sentiment analysis offers these organizations the ability to monitor the different social media sites in real time and act accordingly. Marketing managers, PR firms, campaign managers, politicians, and even equity investors and online shoppers are the direct beneficiaries of sentiment analysis technology.

INTRODUCTION TO SENTIMENT ANALYSIS

Based on slides from Bing Liu

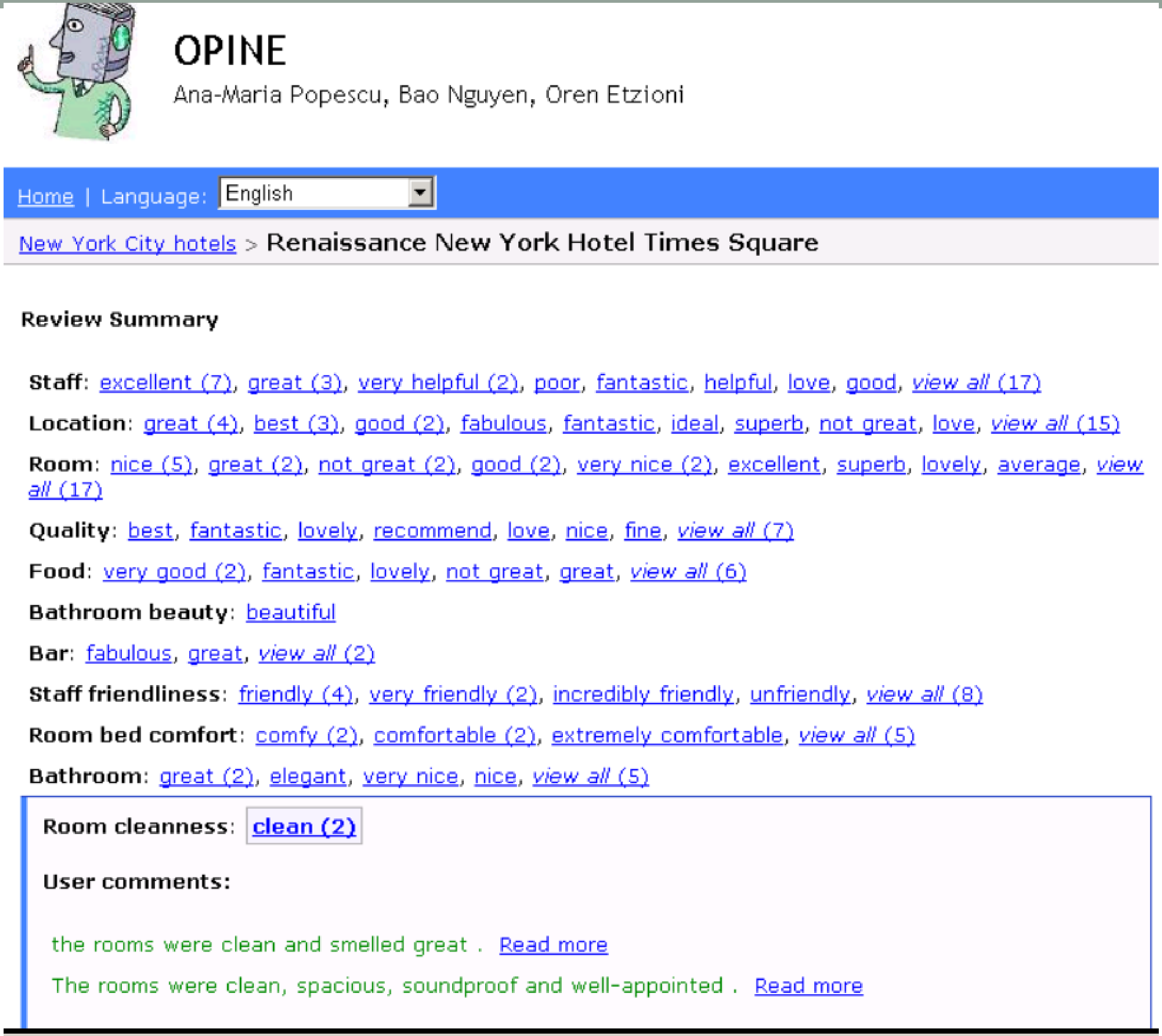
and some of our work

Introduction

- Sentiment analysis
  - Computational study of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in text.
    - Text = Reviews, blogs, discussions, news, comments, feedback ….
- Sometimes called *opinion mining*

Typical sentiment analysis usage

- Extract from text how people feel about different products
- Sentiment analysis can be tricky
  - `Honda Accords and Toyota Camrys are nice sedans`
  - `Honda Accords and Toyota Camrys are nice sedans, but hardly the best car on the road`



OPINE

Ana-Maria Popescu, Bao Nguyen, Oren Etzioni

New York City hotels > Renaissance New York Hotel Times Square

Review Summary

Staff: excellent (7), great (3), very helpful (2), poor, fantastic, helpful, love, good, *view all* (17)

Location: great (4), best (3), good (2), fabulous, fantastic, ideal, superb, not great, love, *view all* (15)

Room: nice (5), great (2), not great (2), good (2), very nice (2), excellent, superb, lovely, average, *view all* (17)

Quality: best, fantastic, lovely, recommend, love, nice, fine, *view all* (7)

Food: very good (2), fantastic, lovely, not great, great, *view all* (6)

Bathroom beauty: beautiful

Bar: fabulous, great, *view all* (2)

Staff friendliness: friendly (4), very friendly (2), incredibly friendly, unfriendly, *view all* (8)

Room bed comfort: comfy (2), comfortable (2), extremely comfortable, *view all* (5)

Bathroom: great (2), elegant, very nice, nice, *view all* (5)

Room cleanness: **clean (2)**

User comments:

the rooms were clean and smelled great . Read more

The rooms were clean, spacious, soundproof and well-appointed . Read more

Sentiment Analysis is Hot

- Dozens of companies and over a thousand research papers

digital trowel
Pure Intelligence.

BUZZLOGIC

LingPipe

Nielsen
BuzzMetrics

umbria

REUTERS

IVolatility.com®

reputica

Opinions are widely stated

- Organization internal data
  - Customer feedback from emails, call centers, etc.
- News and reports
  - Opinions in news articles and commentaries
- Word-of-mouth on the Web
  - Personal experiences and opinions about anything in reviews, forums, blogs, Twitter, micro-blogs, etc
  - Comments about articles, issues, topics, reviews, etc.
  - Postings at social networking sites, e.g., Facebook.

Sentiment analysis applications

- Businesses and organizations
  - Benchmark products and services; market intelligence.
    - Businesses spend a huge amount of money to find consumer opinions using consultants, surveys and focus groups, etc
- Individuals
  - Make decisions to purchase products or to use services
  - Find public opinions about political candidates and issues
- Ad placement: e.g. in social media
  - Place an ad if one praises a product.
  - Place an ad from a competitor if one criticizes a product.
- Opinion retrieval: provide general search for opinions.

Roadmap

- **Sentiment Analysis Problem**
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Opinion summarization
- Sentiment lexicon generation
- Mining comparative opinions
- Some complications
- Generalized sentiment analysis

Problem Statement: Abstraction

- It consists of two parts

(1) Opinion definition. What is an opinion?

(2) Opinion summarization

- Opinions are subjective. An opinion from a single person (unless a VIP) is often not sufficient for action.
- We need opinions from many people, and thus opinion summarization.

Abstraction (1): what is an opinion?

- **Id: Abc123 on 5-1-2008** "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"
- One can look at this review/blog at the
  - document level, i.e., is this review + or -?
  - sentence level, i.e., is each sentence + or -?
  - entity and feature/aspect level

Entity and aspect/feature level

- **Id: Abc123 on 5-1-2008** "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"
- What do we see?
  - Opinion targets: entities and their features/aspects
  - Sentiments: positive and negative
  - Opinion holders: persons who hold the opinions
  - Time: when opinions are expressed

Two main types of opinions

(Jindal and Liu 2006; Liu, 2010)

- Regular opinions: Sentiment/opinion expressions on some target entities
  - Direct opinions:
    - "The touch screen is really cool."
  - Indirect opinions:
    - "After taking the drug, my pain has gone."
- Comparative opinions: Comparisons of more than one entity.
  - E.g., "iPhone is better than Blackberry."
- We focus on regular opinions first, and just call them opinions.

A (regular) opinion

- An *opinion has the following basic components*

  $(g_i, so_{ijkl}, h_i, t_l)$,

  where

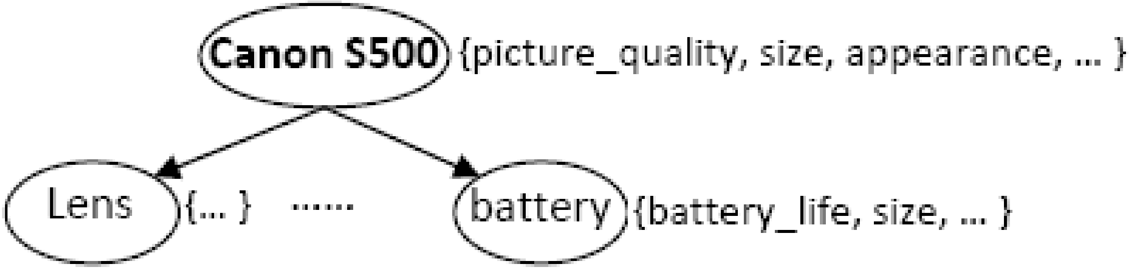
  - g_j is a target
  - so_{ijl} is the sentiment value of the opinion from opinion holder h_i on target g_j at time t_l. so_{ijl} is positive, negative or neutral, or a rating score
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Opinion target

- In some cases, opinion target is a single entity or topic.
  - *"I love iPhone"* and *"I support tax cut."*
- But in many other cases, it is more complex.
  - "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool.*"
    - Opinion target of the 3rd sentence is not just touch screen, but the "touch screen of iPhone".
  - "*I support tax cut for the middle class, but for the rich…*"
- We decompose the opinion target

Entity and aspect (Hu and Liu, 2004; Liu, 2006)

- **Definition (entity)**: An *entity e* is a product, person, event, organization, or topic. *e* is represented as
  - a hierarchy of components, sub-components, and so on.
  - Each node represents a component and is associated with a set of attributes of the component.


Canon S500
{picture_quality, size, appearance, ... }
Lens
{... }
......
battery
{battery_life, size, ... }

- An opinion can be expressed on any node or attribute of the node.
- For simplicity, we use the term ***aspects*** (features) to represent both components and attributes.

Opinion definition (Liu, Ch. in NLP handbook, 2010)

- An *opinion* is a quintuple

  $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,

  where
  - e_j is a target entity.
  - a_{jk} is an aspect/feature of the entity e_j.
  - so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on aspect a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Some remarks about the definition

- Although introduced using a product review, the definition is generic
  - Applicable to other domains,
  - E.g., politics, social events, services, topics, etc.
- (e_j, a_{jk}) is also called the opinion target
  - Opinion without knowing the target is of limited use.
- The five components in (e_j, a_{jk}, so_{ijkl}, h_i, t_l) must correspond to one another. Very hard to achieve
- The five components are essential. Without any of them, it is problematic in general.

Our example blog in quintuples

- **Id: Abc123 on 5-1-2008** "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"
- In quintuples

  (iPhone, GENERAL, +, Abc123, 5-1-2008)

  (iPhone, touch_screen, +, Abc123, 5-1-2008)

  ….

  - We will discuss comparative opinions later.

“Confusing” terminologies

- Entity is also called object.
- Aspect is also called feature, attribute, facet, etc
- Opinion holder is also called opinion source

Reader’s standing point

- See this sentence
  - “I am so happy that Google price shot up today.”
- Although the sentence gives an explicit sentiment, different readers may feel very differently.
  - If a reader sold his Google shares yesterday, he will not be that happy.
  - If a reader bought a lot of Google shares yesterday, he will be very happy.
- Current research mostly ignores the issue.

Structure the unstructured

- Goal: Given an opinionated document,
  - Discover all quintuples $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,
  - Or, solve some simpler form of the problem
    - E.g. classify the sentiment of the entire document
- With the quintuples,
  - Unstructured Text → Structured Data
    - Traditional data and visualization tools can be used to slice, dice and visualize the results.
    - Enables qualitative and quantitative analysis.

Subjectivity

- **Sentence subjectivity:** An *objective sentence* presents some factual information, while a *subjective sentence* expresses some personal opinions, beliefs, views, feelings, or emotions.
  - Not the same as emotion

Subjectivity

- Subjective expressions come in many forms, e.g., opinions, allegations, desires, beliefs, suspicions, and speculations (Wiebe, 2000; Riloff et al 2005).
  - A subjective sentence may contain a positive or negative opinion
- Most opinionated sentences are subjective, but objective (factual) sentences can imply opinions too (Liu, 2010)
  - “The machine stopped working in the second day”
  - “We brought the mattress yesterday, and a body impression has formed.”
  - “After taking the drug, there is no more pain”

Rational and emotional evaluations

- Rational evaluation: Many evaluation/opinion sentences express no emotion
  - e.g., “The voice of this phone is clear”
- Emotional evaluation
  - e.g., “I love this phone”
  - “The voice of this phone is crystal clear” (?)
- Some emotion sentences express no (positive or negative) opinion/sentiment
  - e.g., “I am so surprised to see you”.

Abstraction (2): opinion summary

- With a lot of opinions, a summary is necessary.
  - A multi-document summarization task
- For factual texts, summarization is to select the most important facts and present them in a sensible order while avoiding repetition
  - 1 fact = any number of the same fact
- But for opinion documents, it is different because opinions have a quantitative side & have targets
  - 1 opinion ≠ a number of opinions
  - Aspect-based summary is more suitable
    - Quintuples form the basis for opinion summarization

Feature-based opinion summary[1]

(Hu & Liu, 2004)

"*"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"*

1.

….

Feature Based Summary of iPhone:

Feature1: **Touch screen**

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

…

Feature2: **voice quality**

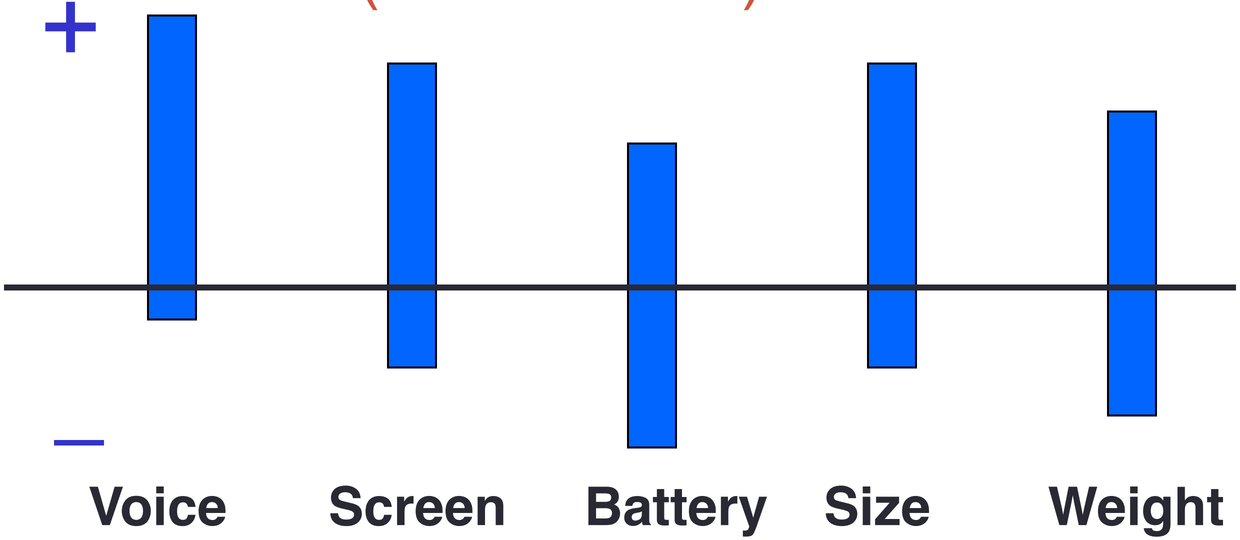
…

Note: We omit opinion holders

Opinion Observer (Liu et al. 2005)

Summary of reviews of

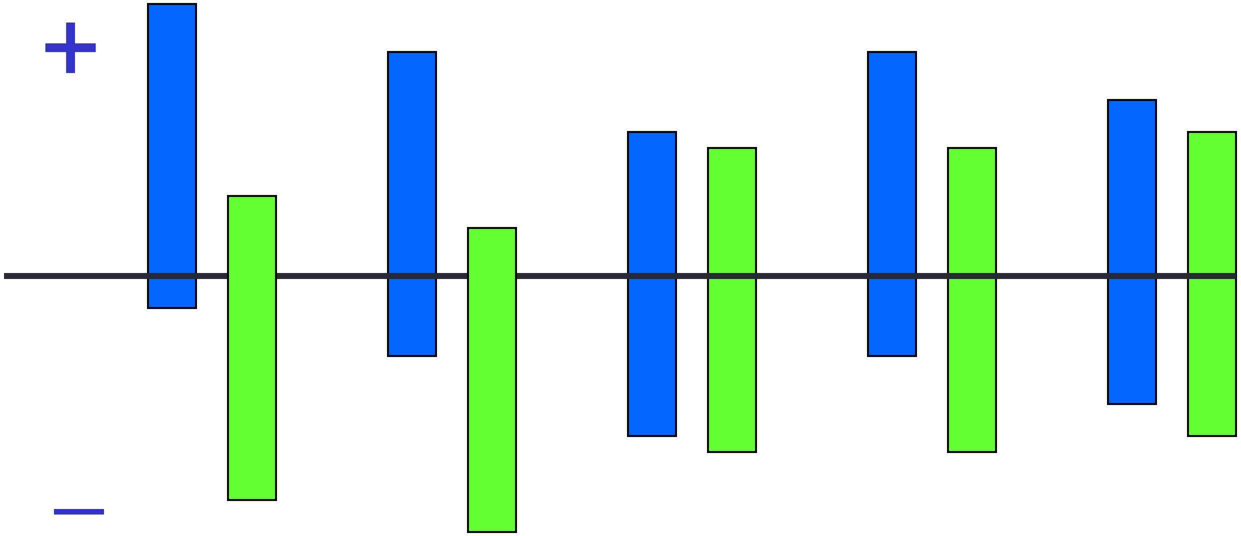
🟦 Cell Phone 1



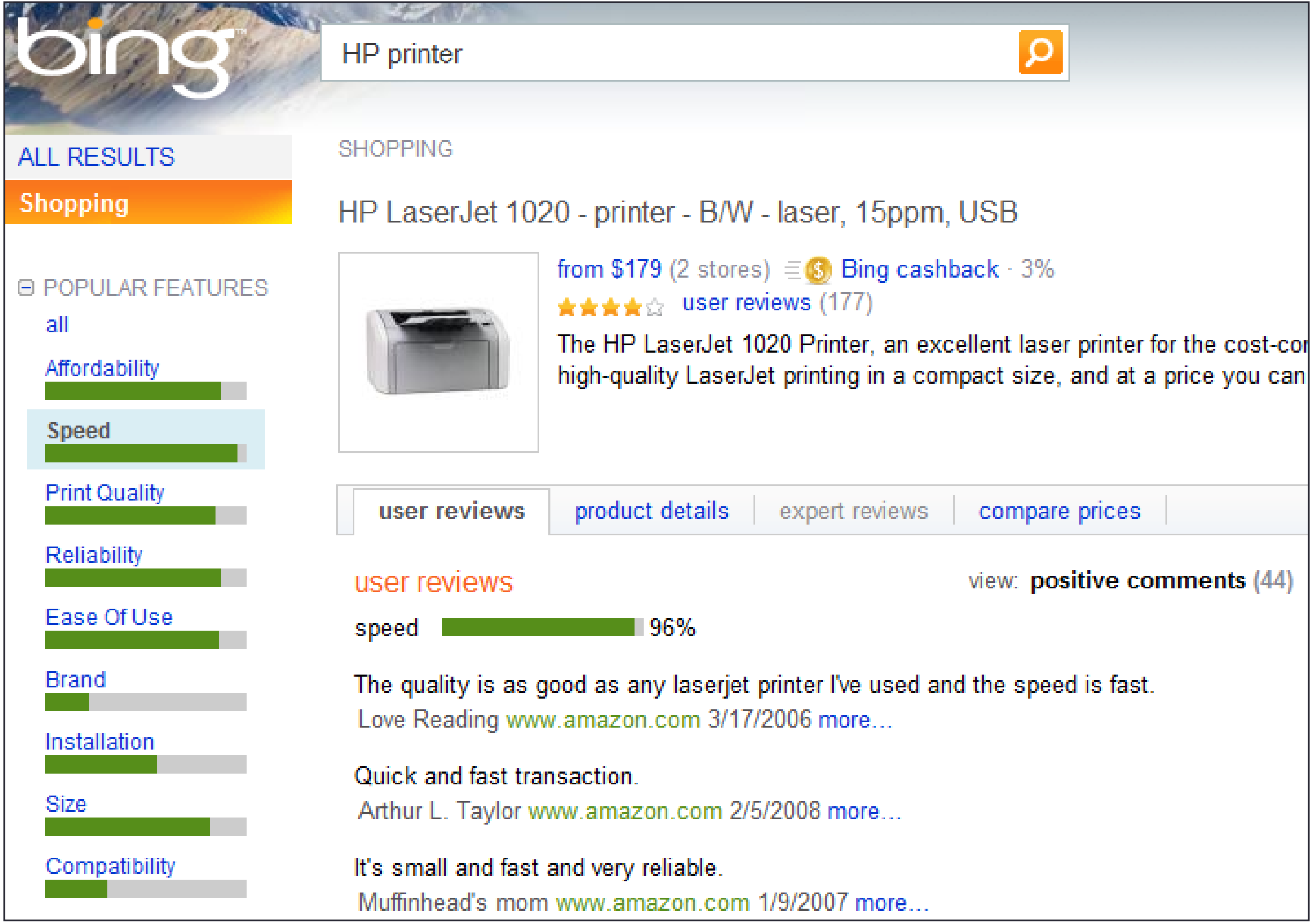
Comparison of reviews of

🟦 Cell Phone 1

🟩 Cell Phone 2



Aspect-based opinion summary



HP printer

ALL RESULTS

Shopping

⊟ POPULAR FEATURES

all

Affordability

Speed

Print Quality

Reliability

Ease Of Use

Brand

Installation

Size

Compatibility

SHOPPING

HP LaserJet 1020 - printer - B/W - laser, 15ppm, USB

from $179 (2 stores) Bing cashback · 3%

★★★★☆ user reviews (177)

The HP LaserJet 1020 Printer, an excellent laser printer for the cost-co
high-quality LaserJet printing in a compact size, and at a price you can

user reviews | product details | expert reviews | compare prices

user reviews

view: **positive comments** (44)

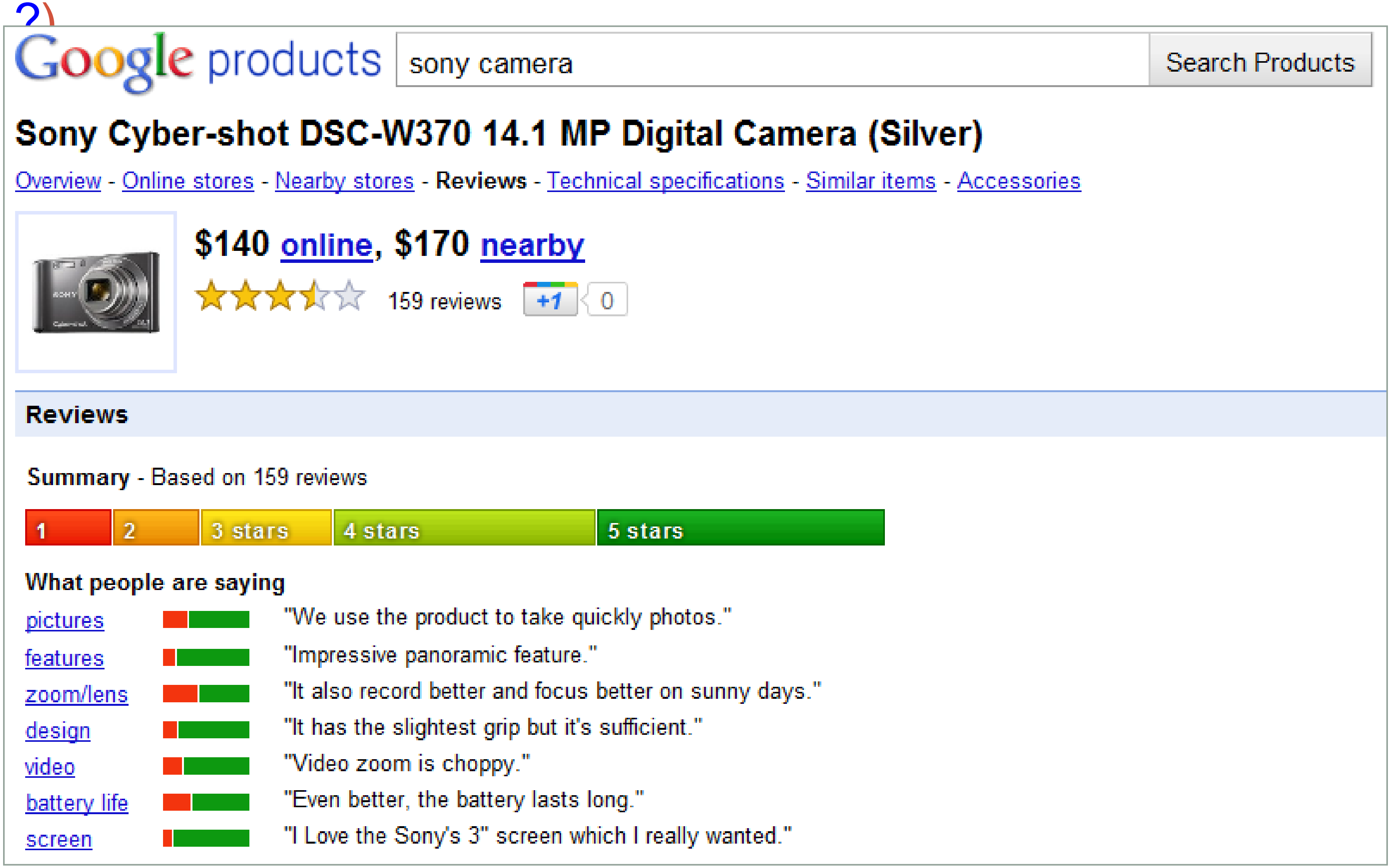
speed 96%

The quality is as good as any laserjet printer I've used and the speed is fast.
Love Reading www.amazon.com 3/17/2006 more...

Quick and fast transaction.
Arthur L. Taylor www.amazon.com 2/5/2008 more...

It's small and fast and very reliable.
Muffinhead's mom www.amazon.com 1/9/2007 more...

Google Product Search (Blair-Goldensohn et al 2008 ?)



Sony Cyber-shot DSC-W370 14.1 MP Digital Camera (Silver)

Overview - Online stores - Nearby stores - **Reviews** - Technical specifications - Similar items - Accessories

$140 online, $170 nearby

★★★½☆ 159 reviews +1 0

Reviews

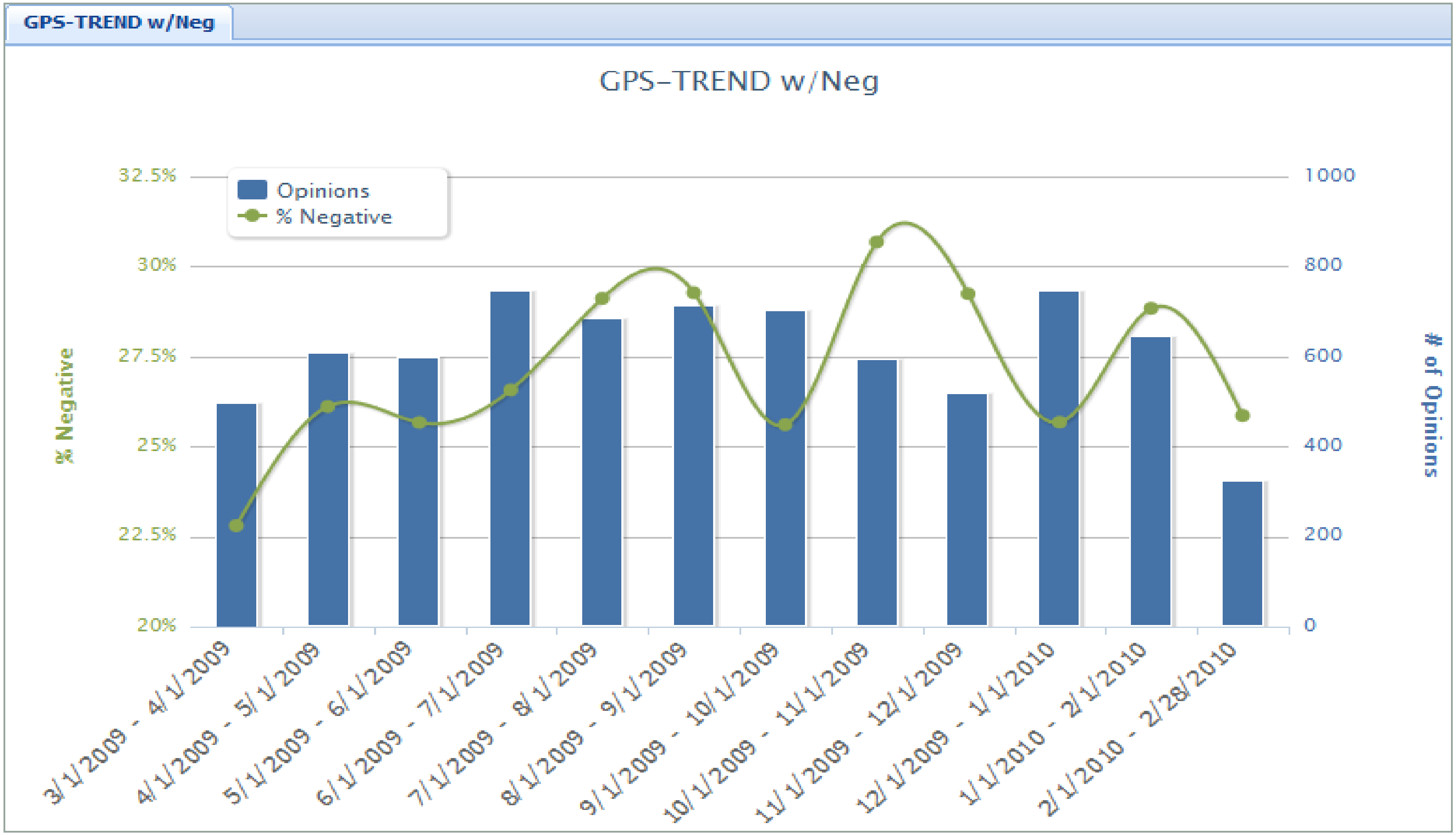
Summary - Based on 159 reviews

1 | 2 | 3 stars | 4 stars | 5 stars

What people are saying

pictures	🟥🟩	"We use the product to take quickly photos."
features	🟥🟩	"Impressive panoramic feature."
zoom/lens	🟥🟩	"It also record better and focus better on sunny days."
design	🟥🟩	"It has the slightest grip but it's sufficient."
video	🟥🟩	"Video zoom is choppy."
battery life	🟥🟩	"Even better, the battery lasts long."
screen	🟥🟩	"I Love the Sony's 3" screen which I really wanted."

Aggregate opinion trend



Sentiment mining requires solving coupled IE problems

- (e_j, a_{jk}, so_{ijkl}, h_i, t_l),
  - e_j - a target entity: Named Entity Extraction (more)
  - a_{jk} – an aspect/feature of e_j: Information Extraction
  - so_{ijkl} is sentiment: Sentiment Identification
  - h_i is an opinion holder: Information/Data Extraction
  - t_l is the time: Information/Data Extraction
  - 5 pieces of information must match
- Coreference and entity resolution

Easier and harder problems

- Tweets from Twitter are probably the easiest
  - short and thus usually straight to the point
  - Stocktwits are much harder! (more on that later)
- Reviews are next
  - entities are given (almost) and there is little noise
- Discussions, comments, and blogs are hard.
  - Multiple entities, comparisons, noisy, sarcasm, etc
- Extracting entities and aspects, and determining sentiments/opinions about them are hard.
- Combining them is harder.

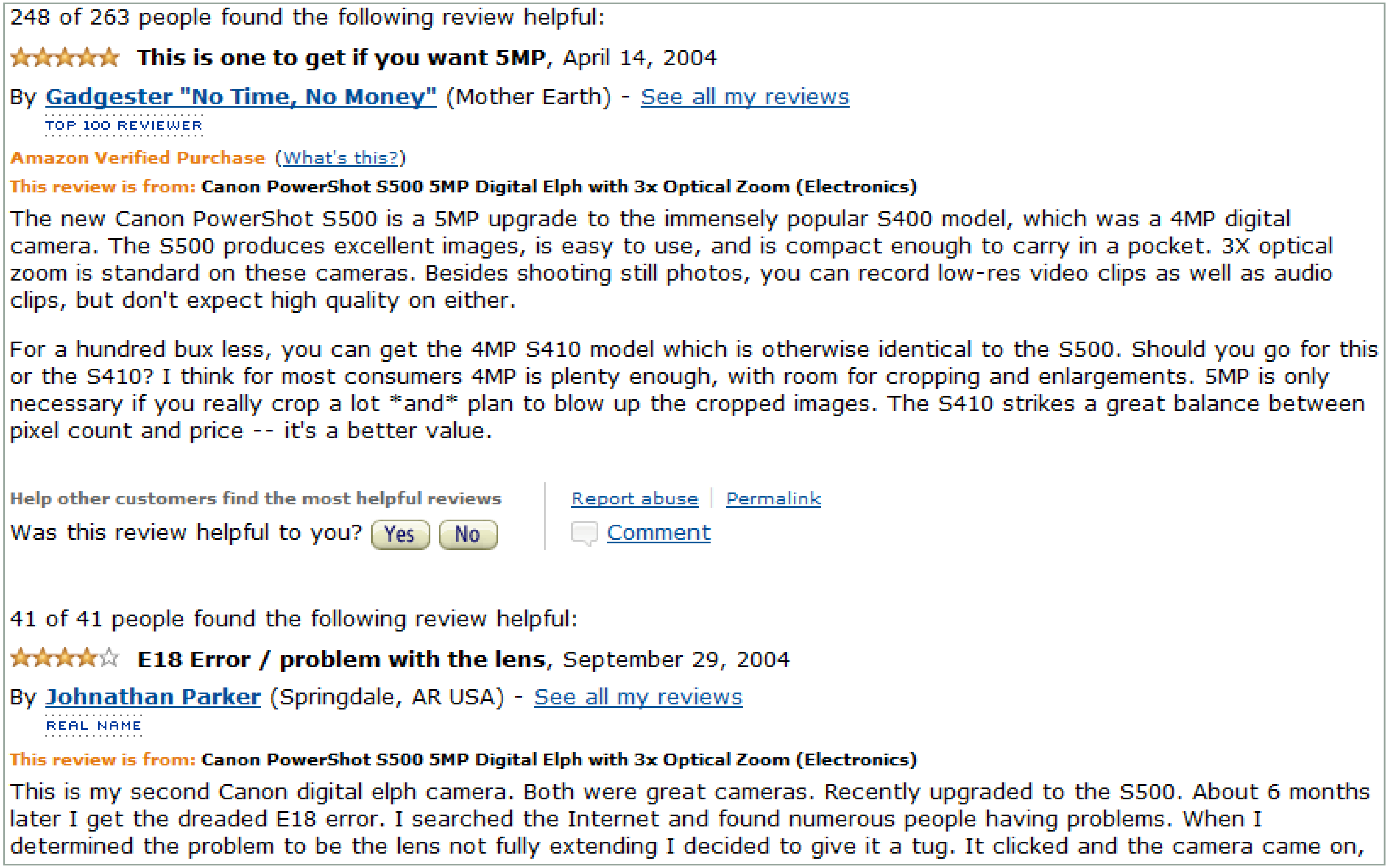
Roadmap

- Sentiment Analysis Problem
- **Document sentiment classification**
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Opinion summarization
- Sentiment lexicon generation
- Mining comparative opinions
- Some complications
- Generalized sentiment analysis

Sentiment classification

- Classify a whole opinion document (e.g., a review) based on the overall sentiment of the opinion holder (Pang et al 2002; Turney 2002)
  - Classes: Positive, negative (possibly neutral)
  - Neutral or no opinion is hard. Most papers ignore it.
- An example review:
  - *"I bought an iPhone a few days ago. It is such a nice phone, although a little large. The touch screen is cool. The voice quality is clear too. I simply love it!"*
  - Classification: positive or negative?
- Perhaps the most widely studied problem.

Some Amazon reviews



248 of 263 people found the following review helpful:

★★★★★ **This is one to get if you want 5MP**, April 14, 2004

By **Gadgester "No Time, No Money"** (Mother Earth) - See all my reviews

TOP 100 REVIEWER

Amazon Verified Purchase (What's this?)

This review is from: Canon PowerShot S500 5MP Digital Elph with 3x Optical Zoom (Electronics)

The new Canon PowerShot S500 is a 5MP upgrade to the immensely popular S400 model, which was a 4MP digital camera. The S500 produces excellent images, is easy to use, and is compact enough to carry in a pocket. 3X optical zoom is standard on these cameras. Besides shooting still photos, you can record low-res video clips as well as audio clips, but don't expect high quality on either.

For a hundred bux less, you can get the 4MP S410 model which is otherwise identical to the S500. Should you go for this or the S410? I think for most consumers 4MP is plenty enough, with room for cropping and enlargements. 5MP is only necessary if you really crop a lot *and* plan to blow up the cropped images. The S410 strikes a great balance between pixel count and price -- it's a better value.

Help other customers find the most helpful reviews

Was this review helpful to you?

41 of 41 people found the following review helpful:

★★★★☆ **E18 Error / problem with the lens**, September 29, 2004

By **Johnathan Parker** (Springdale, AR USA) - See all my reviews

REAL NAME

This review is from: Canon PowerShot S500 5MP Digital Elph with 3x Optical Zoom (Electronics)

This is my second Canon digital elph camera. Both were great cameras. Recently upgraded to the S500. About 6 months later I get the dreaded E18 error. I searched the Internet and found numerous people having problems. When I determined the problem to be the lens not fully extending I decided to give it a tug. It clicked and the camera came on,

A text classification task

- It is basically a text classification problem
- But different from topic-based text classification.
  - In topic-based text classification (e.g., computer, sport, science), topic words are important.
  - In sentiment classification, opinion/sentiment words are more important, e.g., great, excellent, horrible, bad, worst, etc.

Assumption and goal

- Assumption: The doc is written by a single person and express opinion/sentiment on a single entity.
- Goal: discover (_, _, *so*, _, _),

  where e, a, h, and t are ignored
- Reviews usually satisfy the assumption.
  - Almost all papers use reviews
  - Positive: 4 or 5 stars, negative: 1 or 2 stars
- Many forum postings and blogs do not
  - They can mention and compare multiple entities
  - Many such postings express no sentiments

Supervised learning example

- Training and test data
  - Movie reviews with star ratings
    - 4-5 stars as positive
    - 1-2 stars as negative
- Neutral is ignored.
- SVM gives the best classification accuracy based on balanced training data
  - Typical result: 80-90% accuracy
  - Features: unigrams (bag of individual words)

Features for supervised learning

- The problem has been studied by numerous researchers
    - Including domain adaption and cross-lingual, etc.
- Key: feature engineering. A large set of features have been tried by researchers. E.g.,
    - Terms frequency and different IR weighting schemes
    - Part of speech (POS) tags
    - Opinion words and phrases
    - Negations
    - Syntactic dependency

Roadmap

- Sentiment Analysis Problem
- Document sentiment classification
- **Sentence subjectivity & sentiment classification**
- Aspect-based sentiment analysis
- Opinion summarization
- Sentiment lexicon generation
- Mining comparative opinions
- Some complications
- Generalized sentiment analysis

Subjectivity classification

- Document-level sentiment classification is too coarse for most applications.
- So do sentence level analysis
  - Assumes a single sentiment per sentence
  - not always true, so one can classify clauses instead

Sentence sentiment analysis

- Usually consists of two steps
  - Subjectivity classification
    - To identify subjective sentences
  - Sentiment classification of subjective sentences
    - As positive or negative
- But bear in mind
  - Many objective sentences can imply sentiments
  - Many subjective sentences do not express positive or negative sentiments/opinions
    - E.g.,”I believe he went home yesterday.”

Subjectivity classification using patterns

(Rilloff and Wiebe, 2003)

- A bootstrapping approach.
  - A high precision classifier is first used to automatically identify some subjective and objective sentences.
    - Two high precision (but low recall) classifiers are used,
      - A high precision subjective classifier
      - A high precision objective classifier
      - Based on manually collected lexical items, single words and n-grams, which are good subjective clues.
  - A set of patterns are then learned from these identified subjective and objective sentences.
    - Syntactic templates are provided to restrict the kinds of patterns to be discovered, e.g., <subj> passive-verb.
  - The learned patterns are then used to extract more subjective and objective sentences (the process can be repeated).

Roadmap

- Sentiment Analysis Problem
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- **Feature/Aspect-based sentiment analysis**
- Sentiment lexicon generation
- Mining comparative opinions
- Some complications
- Generalized sentiment analysis

We need to go further

- Sentiment classification at both the document and sentence (or clause) levels are useful, but
  - They do not find what people liked and disliked.
- They do not identify the targets of opinions, i.e.,
  - Entities and their aspects
  - Without knowing targets, opinions are of limited use.
- We need to go to the entity and aspect level.

Recall an opinion is a quintuple

- An *opinion* is a quintuple

  $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,

  where

  - e_j is a target entity.
  - a_{jk} is an aspect/feature of the entity e_j.
  - so_{ijkl} is the sentiment value of the opinion of the opinion holder h_i on feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Aspect-based sentiment analysis

- Much of the research is based on online reviews
- For reviews, aspect-based sentiment analysis is easier because the entity (i.e., product name) is usually known
  - Reviewers simply express positive and negative opinions on different aspects of the entity.
- For blogs, forum discussions, etc., it is harder:
  - both entity and aspects of entity are unknown,
  - there may also be many comparisons, and
  - there is also a lot of irrelevant information.

Find entities (entity set expansion)

- Although similar, it is somewhat different from the traditional named entity recognition (NER).
- E.g., one wants to study opinions on phones
  - given Motorola and Nokia, find all phone brands and models in a corpus, e.g., Samsung, Moto,

Feature/Aspect extraction

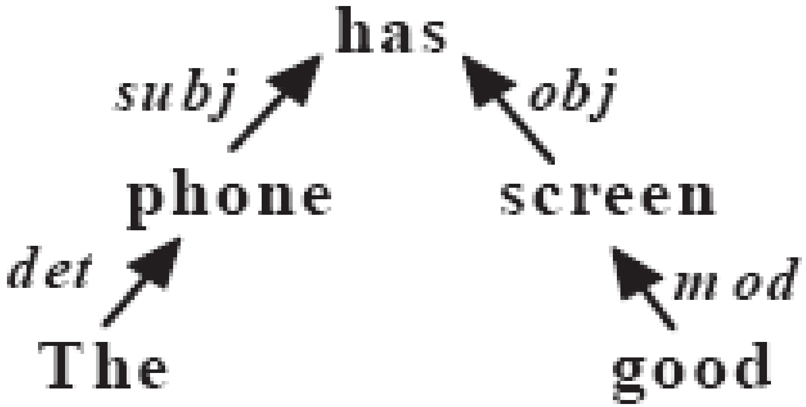
- Extraction may use:
  - frequent nouns and noun phrases
    - Sometimes limited to a set known to be related to the entity of interest or using **part discriminators**
      - e.g., for a scanner entiity"of scanner", "scanner has",
  - opinion and target relations
    - Proximity or syntactic dependency
  - Standard IE methods
    - Rule-based or supervised learning
      - Often HMMs or CRFs (like standard IE)

Extract aspects using DP (Qiu et al. 2009; 2011)

- Use d*ouble propagation* (DP)
  - Like co-training
- an opinion should have a target, entity or aspect.
- DP extracts both aspects and opinion words.
  - Knowing one helps find the other.
  - E.g., "*The rooms are spacious*"

The DP method

- DP is a bootstrapping method
  - Input: a set of seed opinion words,
  - no aspect seeds needed
- Based on dependency grammar (Tesniere 1959).
  - “This phone has good screen”



Roadmap

- Opinion Mining Problem
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- **Sentiment lexicon generation**
- Mining comparative opinions
- Some complications
- Generalized sentiment analysis

Sentiment (or opinion) lexicon

- Sentiment lexicon: lists of words and expressions used to express people's subjective feelings and sentiments/opinions.
  - Not just individual words, but also phrases and idioms, e.g., "cost an arm and a leg"
- Many sentiment lexica can be found on the web
  - They often have thousands of terms, and are quite useful

Sentiment lexicon

- Sentiment words or phrases (also called polar words, opinion bearing words, etc). E.g.,
  - Positive: beautiful, wonderful, good, amazing,
  - Negative: bad, poor, terrible, cost an arm and a leg.
- Many of them are context dependent, not just application domain dependent.
- Three main ways to compile such lists:
  - Manual approach: not a bad idea for a one-time effort
  - Corpus-based approach
  - Dictionary-based approach

Corpus- vs. Dictionary-based method

- Corpus-based approaches
  - Often use a double propagation between opinion words and the items they modify
  - require a large corpus to get good coverage
- Dictionary-based methods
  - Typically use WordNet’s synsets and hierarchies to acquire opinion words
  - usually do not give domain or context dependent meanings

Corpus-based approaches

- Rely on syntactic patterns in large corpora. (Hazivassiloglou and McKeown, 1997; Turney, 2002; Yu and Hazivassiloglou, 2003; Kanayama and Nasukawa, 2006; Ding, Liu and Yu, 2008)
    - Can find domain dependent orientations (positive, negative, or neutral).
- (Turney, 2002) and (Yu and Hazivassiloglou, 2003) are similar.
    - Assign opinion orientations (polarities) to words/phrases.
    - (Yu and Hazivassiloglou, 2003) is slightly different from (Turney, 2002)
        - use more seed words (rather than two) and use log-likelihood ratio (rather than PMI).

Corpus-based approaches (contd)

- Sentiment consistency: Use conventions on connectives to identify opinion words (Hazivassiloglou and McKeown, 1997). E.g.,
  - Conjunction: conjoined adjectives usually have the same orientation.
    - E.g., "This car is *beautiful* **and** *spacious*." (conjunction)
  - AND, OR, BUT, EITHER-OR, and NEITHER-NOR have similar constraints.
  - Learning
    - determine if two conjoined adjectives are of the same or different orientations.
    - Clustering: produce two sets of words: positive and negative

Find domain opinion words

- A similar approach was also taken in (Kanayama and Nasukawa, 2006) but for Japanese words:
  - Instead of only based on intra-sentence sentiment consistency, the new method also looks at the previous and next sentence, i.e., inter-sentence sentiment consistency.
  - Have an initial seed lexicon of positive and negative words.

Context dependent opinion

- Find domain opinion words is insufficient. A word may indicate different opinions in same domain.
    - "The battery life is *long*" (+) and "It takes a *long* time to focus" (-).
- Ding, Liu and Yu (2008) and Ganapathibhotla and Liu (2008) exploited sentiment consistency (both inter and intra sentence) based on contexts
    - It finds context dependent opinions.
    - Context: (adjective, aspect), e.g., (long, battery_life)
    - It assigns an opinion orientation to the pair.
- Other related work (Wu and Wen, 2010; Lu et al., 2011)

The Double Propagation method
(Qiu et al 2009, 2011)

- The DP method can also use dependency of opinions & aspects to extract new opinion words.
- Based on dependency relations
  - Knowing an aspect can find the opinion word that modifies it
    - E.g., “The rooms are spacious”
  - Knowing some opinion words can find more opinion words
    - E.g., “The rooms are spacious and beautiful”
- Jijkoun, Rijke and Weerkamp (2010) did similarly.

Opinions implied by objective terms (Zhang and Liu, 2011)

- Most opinion words are adjectives and adverbs, e.g., good, bad, etc
  - There are also many subjective and opinion verbs and nouns, e.g., hate (VB), love (VB), crap (NN).
- But objective nouns can imply opinions too.
  - E.g., “After sleeping on the mattress for one month, a valley/body impression has formed in the middle.”
- How to discover such nouns in a domain or context?

The technique

- Sentiment analysis to determine whether the context is +ve or –ve.
  - E.g., “I saw a valley in two days, which is terrible.”
  - This is a negative context.
- Statistical test to find +ve and –ve candidates.

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

- Pruning to move those unlikely ones though *sentiment homogeneity*.

Pruning

- For an aspect with an implied opinion, it has a fixed opinion, either +ve or –ve, but not both.
- We find two direct modification relations using a dependency parser.
  - Type 1: $O \rightarrow O\text{-}Dep \rightarrow A$
    - e.g. “ *This TV has good picture quality*.”
  - Type 2: $O \rightarrow O\text{-}Dep \rightarrow H \leftarrow A\text{-}Dep \leftarrow A$
    - e.g. “ *The springs of the mattress are bad*. ”
- If an aspect has mixed opinions based on the two dependency relations, prune it.

Opinions implied by resource usage (Zhang and Liu, 2011)

- Resource usage descriptions may also imply opinions
  - E.g., “This washer uses a lot of water.”
- Two key roles played by resources usage:
  - An important aspect of an entity, e.g., water usage.
  - Imply a positive or negative opinion
- Resource usages that imply opinions can often be described by a triple.

  (verb, quantifier, noun_term),

  - Verb: uses, quantifier: “a lot of “, noun_term: water

Dictionary-based methods

- Typically use WordNet’s synsets and hierarchies to acquire opinion words
  - Start with a small seed set of opinion words.
  - Bootstrap the set to search for synonyms and antonyms in WordNet iteratively (Hu and Liu, 2004; Kim and Hovy, 2004; Valitutti, Strapparava and Stock, 2004; Mohammad, Dunne and Dorr, 2009).
  - Kamps et al., (2004) proposed a WordNet distance method to determine the sentiment orientation of a given adjective.

Semi-supervised learning
(Esuti and Sebastiani, 2005)

- Use supervised learning
  - Given two seed sets: positive set P, negative set N
  - The two seed sets are then expanded using synonym and antonymy relations in an online dictionary to generate the expanded sets P’ and N’.
- P’ and N’ form the training sets.
- Using all the glosses in a dictionary for each term in P’ $\cup$ N’ and converting them to a vector
- Build a binary classifier
  - Tried various learners.

Multiple runs of bootstrapping
(Andreevskaia and Bergler, 2006)

- Basic bootstrapping with given seeds sets (adjectives)
  - First pass: seed sets are expanded using synonym, antonyms, and hyponyms relations in WordNet.
  - Second pass: it goes through all WordNet glosses and identifies the entries that contain in their definitions the sentiment-bearing words from the extended seed set and adds these head words to the corresponding category (+ve, -ve, neutral)
  - Third pass: clean up using a POS tagger to make sure the words are adjectives and remove contradictions.

Multiple runs of bootstrapping (contd)

- Each word is then assigned a fuzzy score reflecting the degree of certainty that the word is opinionated (+ve/-ve).
- The method performs multiple runs of bootstrapping using non-overlapping seed sets.
  - A net overlapping score for each word is computed based on how many times the word is discovered in the runs as +ve (or –ve)
  - The score is normalized based on the fuzzy membership.

Which approach to use?

- Both corpus and dictionary based approaches are needed.
- Dictionary usually does not give domain or context dependent meanings
  - Corpus is needed for that
- Corpus-based approach is hard to find a very large set of opinion words
  - Dictionary is good for that
- In practice, corpus, dictionary and manual approaches are all needed.

Roadmap

- Sentiment Analysis Problem
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Sentiment lexicon generation
- **Mining comparative opinions**
- Some complications
- Generalized sentiment analysis

Comparative Opinions
(Jindal and Liu, 2006)

- *Gradable*
  - *Non-Equal Gradable*: Relations of the type *greater* or *less than*
    - *Ex: “optics of camera A is better than that of camera B”*
  - *Equative*: Relations of the type *equal to*
    - Ex: “*camera A and camera B both come in 7MP*”
  - *Superlative*: Relations of the type *greater* or *less than all others*
    - Ex: “*camera A is the cheapest in market*”

An example

- Consider the comparative sentence
  - “*Canon’s optics is better than those of Sony and Nikon*.”
  - Written by John in 2010.
- The extracted comparative opinion/relation:
  - ({Canon}, {Sony, Nikon}, {optics}, *preferred*:{Canon}, John, 2010)

Common comparatives

- In English, comparatives are usually formed by adding -*er* and superlatives are formed by adding -*est* to their base adjectives and adverbs
- Adjectives and adverbs with two syllables or more and not ending in *y* do not form comparatives or superlatives by adding -*er* or -*est*.
  - Instead, *more*, *most*, *less*, and *least* are used before such words, e.g., *more beautiful*.
- Irregular comparatives and superlatives, i.e., *more most*, *less*, *least*, *better*, *best*, *worse*, *worst*, etc

Roadmap

- Sentiment Analysis Problem
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Sentiment lexicon generation
- Mining comparative opinions
- **Some complications**
- Generalized sentiment analysis

Sentiment shifters (e.g., Polanyi and Zaenen 2004)

- Sentiment/opinion shifters (also called **valence shifters** are words and phrases that can shift or change opinion orientations.
- Negation words like *not*, *never*, *cannot*, etc., are the most common type.
- Many other words and phrases can also alter opinion orientations. E.g., modal auxiliary verbs (e.g., *would*, *should*, *could, etc*)
  - “The brake could be improved.”

Sentiment shifters (contd)

- Some presuppositional items also change opinions, e.g., *barely* and *hardly*
  - “It hardly works.” (comparing to “it works”)
  - It presupposes that better was expected.
- Words like *fail*, *omit*, *neglect* behave similarly,
  - “This camera fails to impress me.”
- Sarcasm changes orientation too
  - “What a great car, it did not start the first day.”
- Jia, Yu and Meng (2009) designed some rules based on parsing to find the scope of negation.

Explicit and implicit aspects
(Hu and Liu 2004)

- Explicit aspects: Aspects explicitly mentioned as nouns or noun phrases in a sentence
  - The picture quality of this phone is great.
- Implicit aspects: Aspects not explicitly mentioned in a sentence but are implied
  - "This car is so expensive."
  - "This phone will not easily fit in a pocket.
  - "Included 16MB is stingy"
- Not much work has been done on mining or mapping implicit aspects.

Implicit aspect mapping

- There are many types of implicit aspect expressions. Adjectives and adverbs are perhaps the most common type.
  - Most adjectives modify or describe specific attributes of entities.
  - "expensive" ⇒ aspect "price," "beautiful" ⇒ aspect "appearance", "heavy" ⇒ aspect "weight"
- Although manual mapping is possible, in different contexts, the meaning can be different.
  - E.g., "The computation is expensive".

Yet more complications

- Spam sentiment
  - Common on user product reviews
- All of NLP
  - Coreference resolution
  - …

Roadmap

- Sentiment Analysis Problem
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Sentiment lexicon generation
- Mining comparative opinions
- Some complications
- **Generalized sentiment analysis**

Beyond positive/negative sentiments

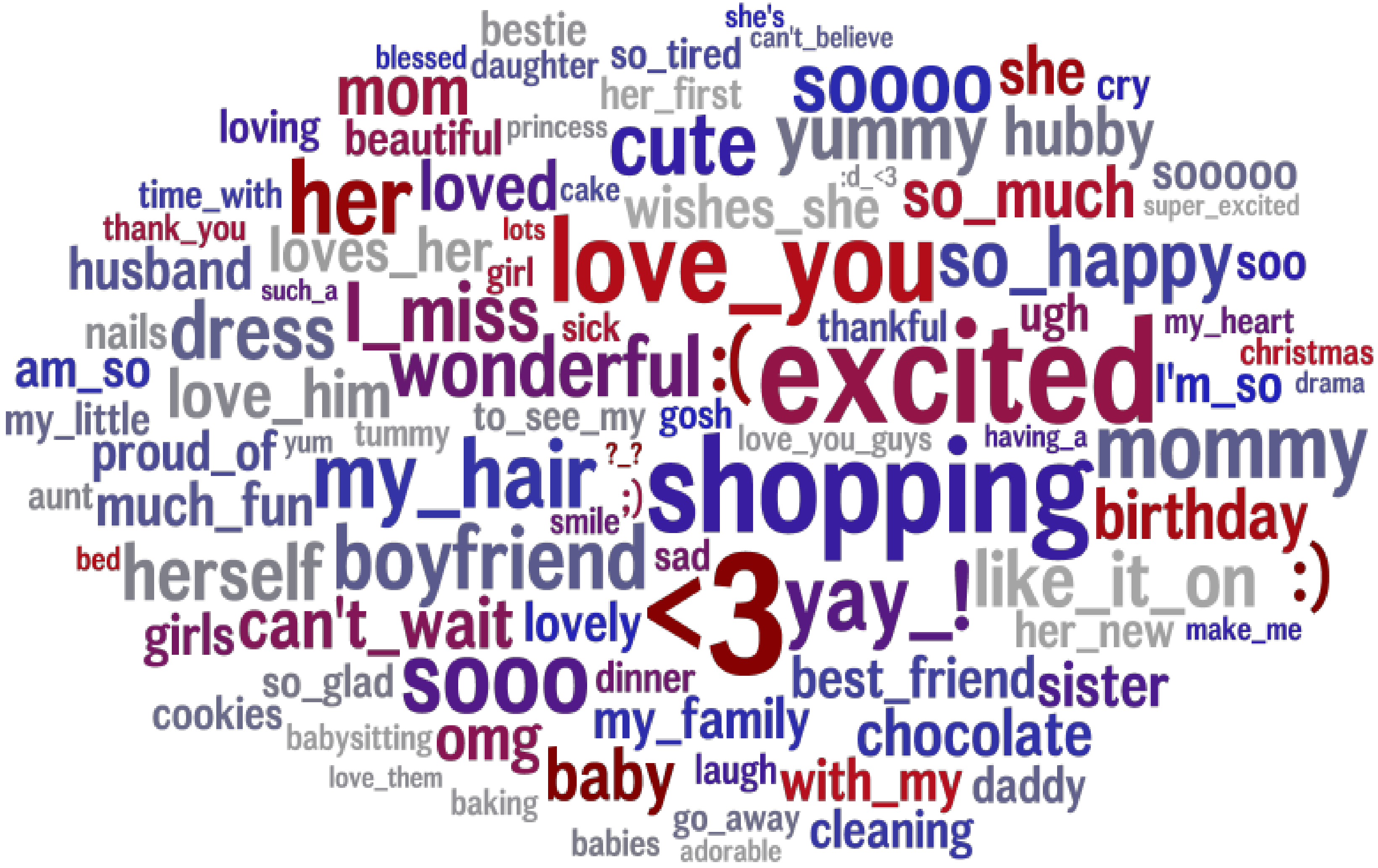
- Why are you buying/reading/tweeting this?
  - What leads to word-of-mouth?
- What emotion does it evoke in you?

Estimating personality and well-being from social media

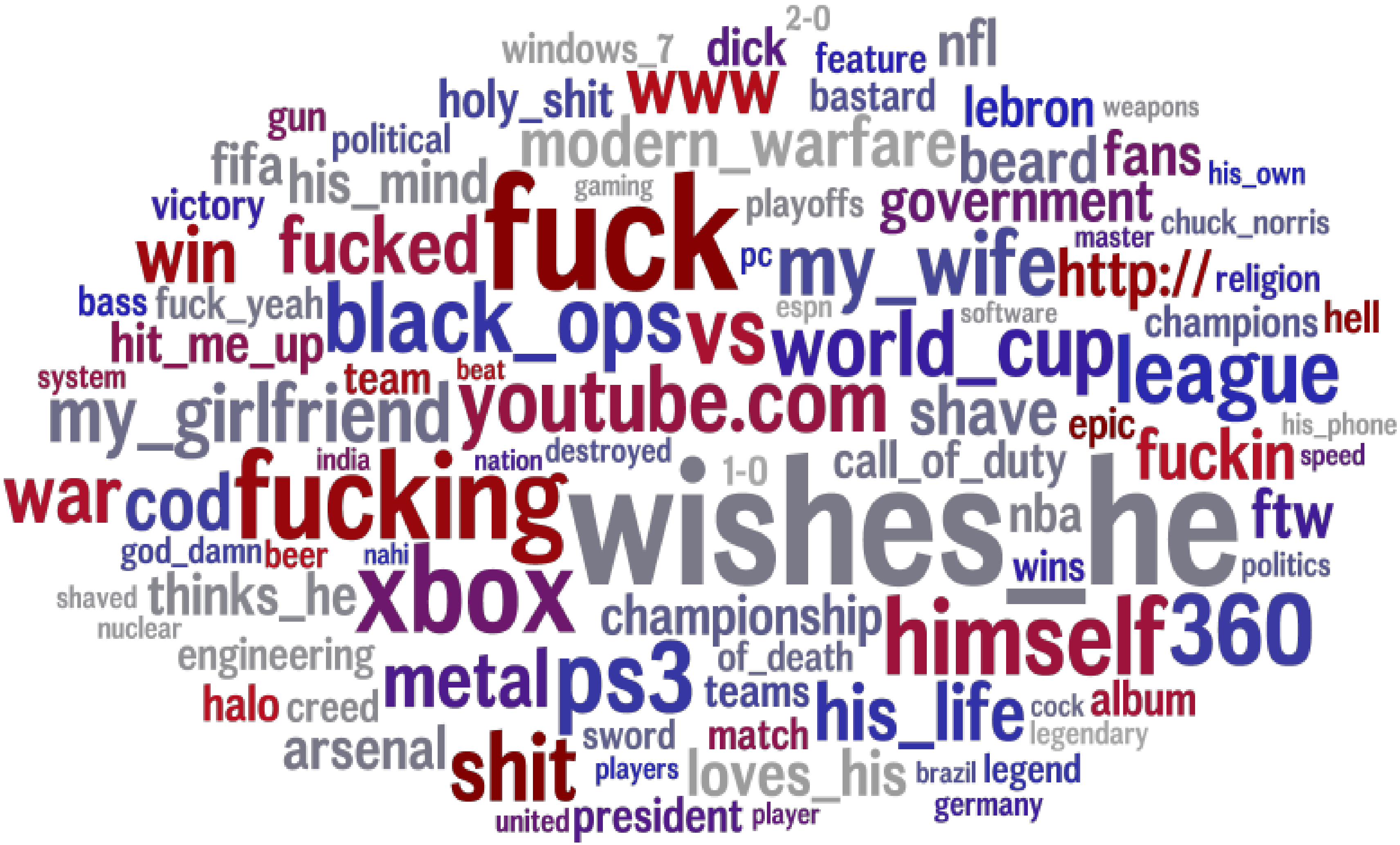
- Correlate words in Facebook posts, tweets and Google search queries with personality, Age, Sex, IQ, happiness
- **Next: results from 100,000 individuals who took the Big Five Personality test and their Facebook posts**
  - Words that most highly correlate with traits

WWBP: Ungar, Schwartz, Johannes, Kern, Ramones, Seligman

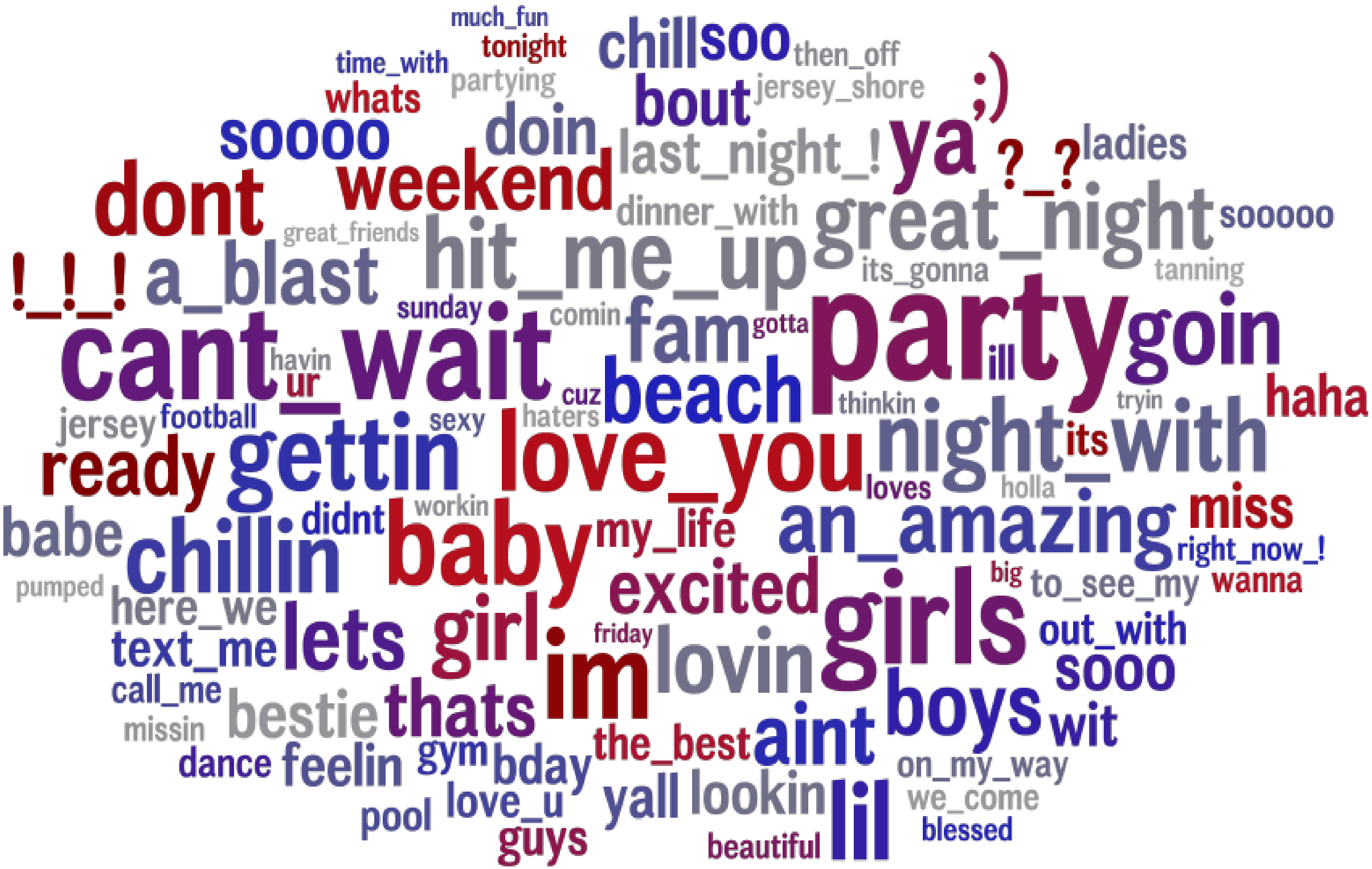
Female


bestie
she's
can't_believe
blessed
daughter
so_tired
mom
her_first
sooooo
she
cry
loving
beautiful
princess
cute
yummy
hubby
time_with
her
loved
cake
:d_<3
sooooo
thank_you
wishes_she
so_much
super_excited
lots
husband
loves_her
girl
love_you
so_happy
soo
such_a
nails
dress
I_miss
sick
thankful
ugh
my_heart
christmas
am_so
love_him
wonderful
:(
excited
I'm_so
drama
my_little
tummy
to_see_my
gosh
love_you_guys
having_a
proud_of
yum
?_?
mommy
aunt
much_fun
my_hair
shopping
smile
;)
birthday
bed
herself
boyfriend
sad
like_it_on
:)
girls
can't_wait
lovely
<3
yay_!
her_new
make_me
so_glad
sooo
dinner
best_friend
sister
cookies
babysitting
omg
my_family
chocolate
love_them
baby
laugh
with_my
daddy
baking
go_away
cleaning
babies
adorable

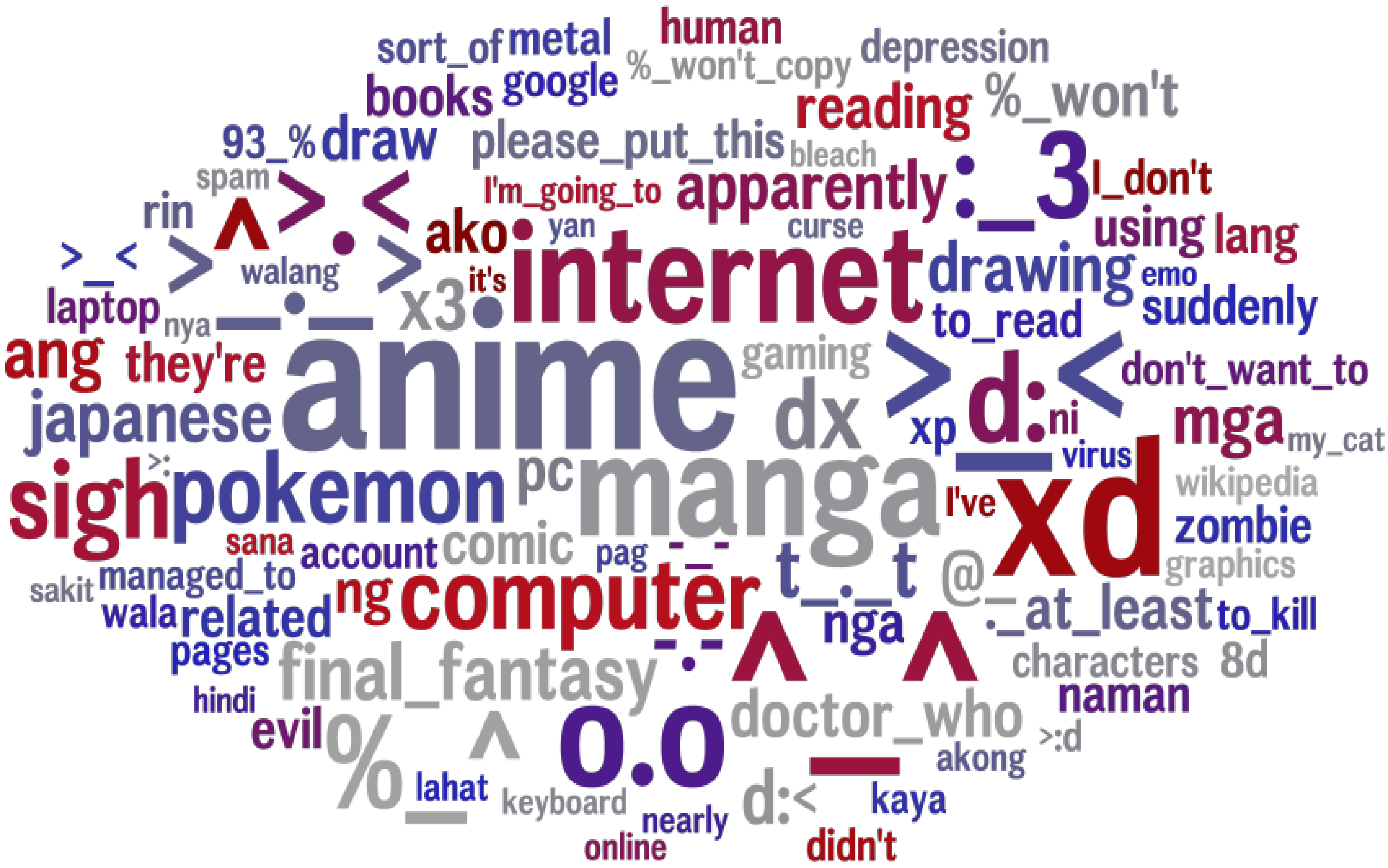
Male


2-0
windows_7
dick
feature
nfl
holy_shit
www
bastard
lebron
weapons
gun
political
modern_warfare
fifa
his_mind
beard
fans
his_own
victory
gaming
playoffs
government
chuck_norris
master
win
fucked
fuck
pc
my_wife
http://
religion
bass
fuck_yeah
black_ops
espn
software
champions
hell
hit_me_up
vs
world_cup
league
system
team
beat
my_girlfriend
youtube.com
shave
epic
his_phone
india
nation
destroyed
1-0
call_of_duty
fuckin
speed
war
cod
fucking
wishes
he
nba
ftw
god_damn
beer
nahi
wins
politics
shaved
thinks_he
xbox
championship
himself
360
nuclear
engineering
of_death
metal
ps3
teams
halo
creed
his_life
cock
album
legendary
arsenal
sword
match
shit
players
loves_his
brazil
legend
united
president
player
germany

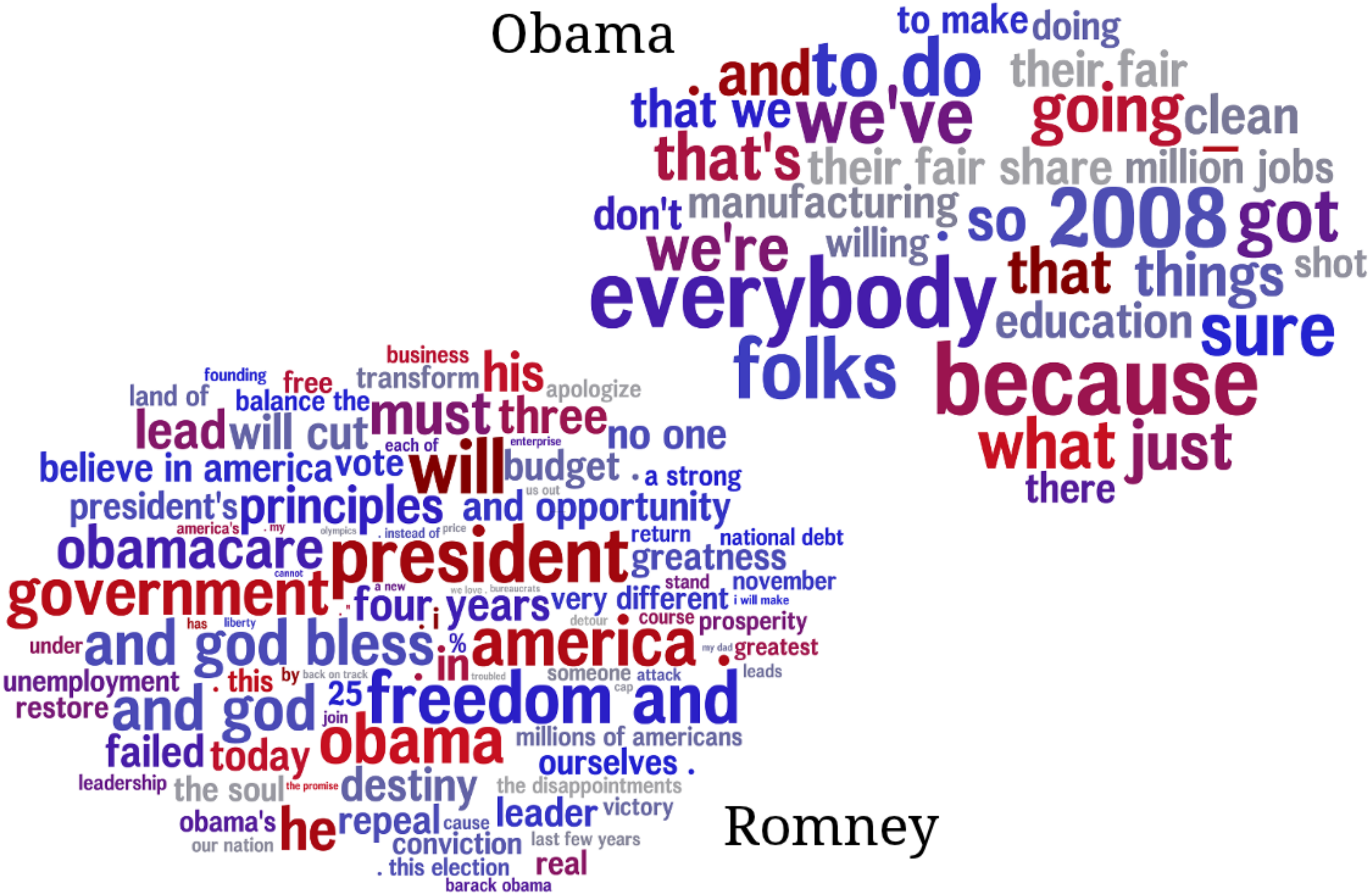
Extroverts


much_fun
tonight
time_with
chill
soo
then_off
partying
whats
bout
jersey_shore
;)
soooo
doin
last_night_!
ya
?_?
ladies
dont
weekend
dinner_with
great_night
sooooo
great_friends
!_!_!
a_blast
hit_me_up
its_gonna
tanning
sunday
comin
fam
gotta
party
goin
cant_wait
havin
ur
ill
cuz
beach
thinkin
tryin
haha
jersey
football
sexy
haters
night
its
with
ready
gettin
love_you
loves
holla
workin
didnt
babe
chillin
baby
my_life
an_amazing
miss
right_now_!
pumped
big
to_see_my
wanna
excited
here_we
friday
girls
out_with
text_me
lets
girl
im
lovin
sooo
call_me
boys
wit
missin
bestie
thats
the_best
aint
dance
feelin
gym
bday
on_my_way
we_come
pool
love_u
yall
lookin
lil
blessed
guys
beautiful

Introverts


sort_of
metal
human
%_won't_copy
depression
books
google
reading
%_won't
93_%
draw
please_put_this
bleach
spam
rin
^
>.<
I'm_going_to
apparently
:_3
I_don't
ako
yan
curse
using
lang
>_<
walang
>_._>
it's
internet
drawing
emo
laptop
nya
x3
to_read
suddenly
ang
they're
anime
gaming
>d:<
don't_want_to
japanese
dx
xp
ni
mga
my_cat
>:
sigh
pokemon
pc
manga
virus
wikipedia
I've
xd
zombie
sana
account
comic
pag
-_-
t_t
@
graphics
sakit
managed_to
wala
related
ng
computer
nga
._.
at_least
to_kill
pages
final_fantasy
-.-
^_^
characters
8d
hindi
naman
evil
%_^
o.o
doctor_who
>:d
akong
lahat
keyboard
nearly
d:<
kaya
online
didn't

Obama vs. Romney


Obama
to make doing
and to do their fair
that we we've going clean
that's their fair share million jobs
don't manufacturing so 2008 got
we're willing that things shot
everybody education sure
folks because
what just
there
business
founding free transform his apologize
land of balance the
lead will cut must three no one
each of enterprise
believe in america vote will budget . a strong
us out
president's principles and opportunity
america's . my olympics . instead of price
return national debt
obamacare president greatness
cannot
government a new we love . bureaucrats stand november
four years very different i will make
has liberty i detour course prosperity
under and god bless % america . my dad greatest
unemployment . this by back on track in troubled someone attack leads
25 cap
restore and god join freedom and
failed today obama millions of americans
ourselves .
leadership the soul the promise destiny the disappointments
obama's he repeal cause leader victory
our nation conviction last few years
. this election real
barack obama
Romney

Sentiment Summary

- Sentiment analysis has many sub-problems.
  - Despite the challenges, applications are flourishing!
- Building a reasonably accurate domain specific system is possible.
  - Building an accurate generic system is very hard.
  - But I am optimistic!

Some interesting sentences

- “Trying out Google chrome because Firefox keeps crashing.”
  - The opinion about Firefox is clearly negative, but for Google chrome, there is no opinion.
  - We need to segment the sentence into clauses to decide that “crashing” only applies to Firefox.
  - “Trying out” also indicates no opinion.
- How about this
  - “I changed to Audi because BMW is so expensive.”

Some interesting sentences (contd)

- Conditional sentences are hard to deal with (Narayanan et al. 2009)
    - “If I can find a good camera, I will buy it.”
    - But conditional sentences can have opinions
        - “If you are looking for a good phone, buy Nokia”
- Questions may or may not have opinions
    - No sentiment
        - “Are there any great perks for employees?”
    - With sentiment
        - “Any idea how to repair this lousy Sony camera?”

Some interesting sentences (contd)

- Sarcastic sentences
  - “What a great car, it stopped working in the second day.”
- Sarcastic sentences are very common in political blogs, comments and discussions.
  - They make political blogs difficult to handle
  - Many political aspects can also be quite complex and hard to extract because they cannot be described using one or two words.
- Some initial work by (Tsur, Davidov, Rappoport 2010)

Some interesting sentences (contd)

- See these two sentences in a medical domain:
  - “I come to see my doctor because of severe pain in my stomach”
  - “After taking the drug, I got severe pain in my stomach”
- If we are interested in opinions on a drug, the first sentence has no opinion, but the second implies negative opinion on the drug.
  - Some understanding seems to be needed?

Some interesting sentences (contd)

- The following two sentences are from reviews in the paint domain.
  - “For paint_X, one coat can cover the wood color.”
  - “For paint_Y, we need three coats to cover the wood color.
- We know that paint_X is good and Paint_Y is not, but how by a system.
  - Do we need commonsense knowledge and understanding of the text?

Some more interesting/hard sentences

- “My goal is to have a high quality tv with decent sound”
- “The top of the picture was much brighter than the bottom.”
- “Google steals ideas from Bing, Bing steals market share from Google.”
- “When I first got the airbed a couple of weeks ago it was wonderful as all new things are, however as the weeks progressed I liked it less and less.”

UNSUPERVISED INFORMATION EXTRACTION

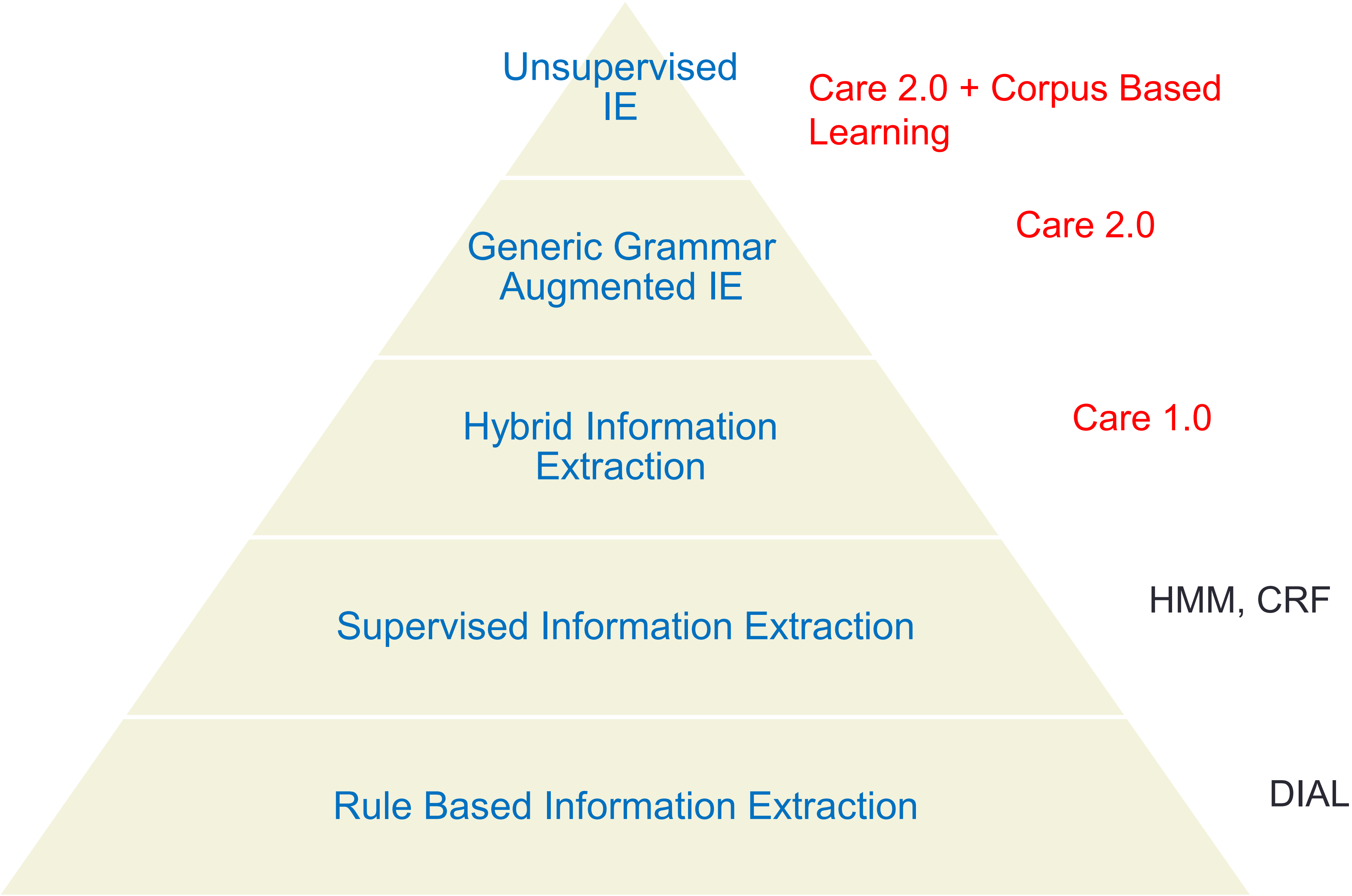
Why do we need Text Mining

- Information Overload
  - News Sites
  - Social Media
  - Scientific Articles
- Text Mining and Information Extraction
  - Extract Key Events
  - Extract Relations
  - Detect Sentiment

Traditional Text Mining is not cost effective nor time efficient

- Why?
  - takes too long to develop
  - too expensive
  - not accurate enough
  - lacks complete coverage

The Evolution of Information Extraction Technology



Unsupervised IE

Care 2.0 + Corpus Based Learning

Generic Grammar Augmented IE

Care 2.0

Hybrid Information Extraction

Care 1.0

Supervised Information Extraction

HMM, CRF

Rule Based Information Extraction

DIAL

Relation extraction

- Relation Extraction (RE) is the task of recognizing instances of specific relationships between two or more entities in a natural language text.
- In a traditional setting, the target relation types are known to a RE system in advance, and it can be prepared for its task either

1. by a knowledge engineer hand-crafting the extraction rules
2. or by the system itself learning the rules from a set of hand-labeled training examples.

- Both ways require a large expenditure of manual labor.

Relation extraction

- In recent years, [Banko and Etzioni 2008] introduced a new setting for the RE task, called Open Information Extraction (Open IE).
- In this setting, the RE system

1. does not know the target relations in advance, and
2. cannot have any relation-specific human input.

- The task requires the system itself to identify the target relations and to train itself for extracting them.

To parse or not to parse?

- IE systems that work on free text can perform either a shallow or a deep parsing.
- The advantages of shallow parsing are high speed, simplicity of training and usage, and consequent higher accuracy.
- Most of the state-of-the-art IE systems do not use deep parsing, because its obvious theoretical advantages are offset by the practical limitations of existing general-purpose parsers.

To parse!!!

- We do perform deep parsing, using a parser which is built specifically for the task of Information Extraction.
- The parser's underlying framework, called CARE-II, is capable of parsing arbitrary weighted typed-feature-structure-based context free grammars.
- This gives the framework a unique power, allowing it to use a high-level unification-based grammar, such as HPSG, while still being able to flexibly interface with feature-rich sequence classifiers, such as CRF-based NER and PoS taggers

Parsing approach

- The parser we built on top of the CARE-II framework is generic – it uses an HPSG-like grammar (derived primarily from the example grammar in [Sag, Wasow et al. 2003]).
- The parser relies on the CRF-trained NER and PoS sequence classifiers to provide weights for different possible typed-feature structure assignments for different words.
- For any input sentence, the parser is able to generate a single highest- weight parse – the parse which is the most consistent with the NER and POS classifiers.

An IE focused parser

- The resulting parses contain many errors, and would probably not compare favorably with the results of the best existing standalone parsers.
- However, the goal of the system is not to provide a general-purpose parser, but to be easily adaptable for the IE task.
- Given a target relation type, the system only requires the definitions for a small number of the relevant content words in order to function as an accurate parser for the *relevant* sentences – the ones that actually contain instances of the target relations.

Sequence classifiers

- A sequence classifier is a component that takes as input a sequence of tokens (the input sentence) and selects for each token a single label from a small predefined set of labels.
- For example:

1. a named entity recognizer (NER) would select the labels from a set like {“Person”, “Organization”, “Location”, “None”}
2. a part-of-speech tagger (PoS) would select the labels from a set like {“NN”, “NNP”, “JJ”, “VB”, “VBG”, …}.

Using the sequence classifiers

- Instead of first running a classifier and then using the produced single fixed sequence of labels, the framework retrieves the actual weights that the classifier assigns, and uses them directly as the weights of the terminal symbols.
- Instead of the classifier's Vitterbi algorithm, it is the more general WTFSCFG inference algorithm, which finds the highest-scoring grammar *parse*, which maximizes the sum of the weights of words, rules, and classifier labels that participate in it.

The actual parsing algorithm

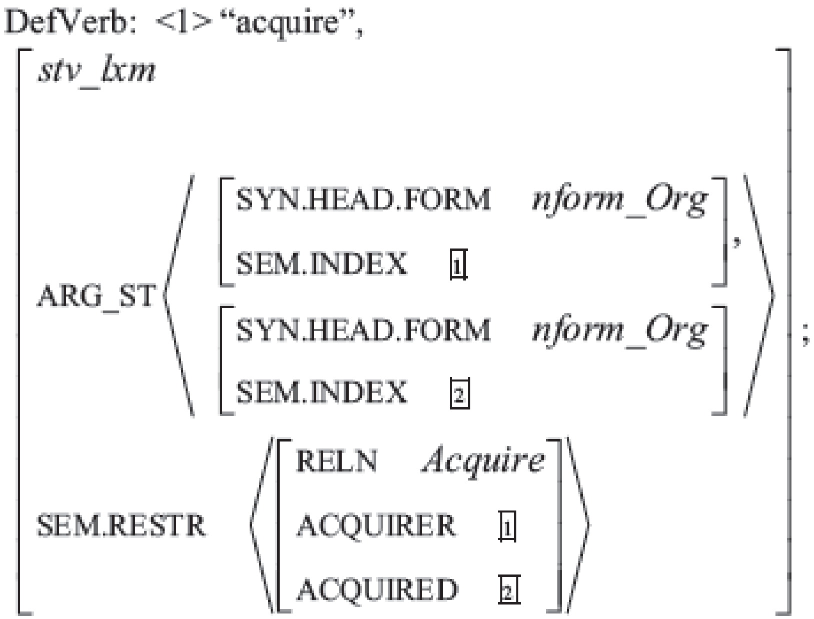
- An extension of the Agenda-based parser for PCFG-s [Klein and Manning 2001]. The extension allows arbitrary weights, as well as feature structures (HPSG like) and their manipulation.
- In the general case, inference with feature structures is Turing-machine-powerful, so the algorithm's worst-case space and time complexity is exponential. However, with careful writing of the grammar rules and feature structures, it can be made to stay efficiently polynomial.

The generic grammar

- The English grammar is based on HPSG, as described in [Sag, Wasow et al. 2003] and adapted for the CARE-II framework. It contains around thirty general rules and the lexicon definitions for several hundred functional words: determiners, pronouns, prepositions, auxiliary verbs, etc.
- The content words are defined generically, with the help of the PoS and NER sequence classifiers. The definitions allow any word in the input sentence to be assigned to any word class (in the HPSG sense), with the weights set as specified by the classifiers

example

- *Qualcomm has acquired Elata for 57 million in cash.*
- *Acquisition*(*Acquirer* = “*Qualcomm*” *Acquired* = “*Elata*”)
- It is sufficient to define a single content word – the verb “to acquire”. The definition can look like this:



DefVerb: <1> “acquire”,

stv_lxm

ARG_ST ⟨ [SYN.HEAD.FORM *nform_Org*, SEM.INDEX 1], [SYN.HEAD.FORM *nform_Org*, SEM.INDEX 2] ⟩

SEM.RESTR ⟨ [RELN *Acquire*, ACQUIRER 1, ACQUIRED 2] ⟩ ;

Example - cont.

- The additional positive weight encourages the parses involving the verb “acquire” over generically defined words.
- If this definition is added to the grammar, the sentence above will be parsed into the following:
- <1: RELN=Org> Qualcomm </1>
- <2: RELN=perfect ARG=3> has </2>
- <3: RELN=Acquire ACQUIRER=1 ACQUIRED=4 PAST=true> acquired </3>
- <4: RELN=Org> Elata </4>

Pattern Learning flow

1. Parse a large unlabeled corpus using the generic grammar. This may result in many parsing mistakes, but genuine meaningful patterns would still appear much more frequently in the parsed text than random mistakes.
2. The extracted frequent patterns are converted into lexical entries and the corpus is reparsed using the combined grammar. This produces much more precise results and also significantly increases recall.
3. The relations extracted by different lexical patterns are filtered, compared to each other, and merged together into single relation types if they have sufficient overlap.
4. Names are given to the relation types and their slots, and the lexical entries are re-generated accordingly, producing the final domain-specific lexicon.

Pattern types

- The patterns are pieces of parses, and the parses are word dependency graphs. Thus, all patterns are connected dependency sub-graphs, and each one includes at least two entity placeholders. There are three sets of pattern types:

1. verb-based
2. noun-based
3. BE-based.

Verb based patterns

- In these patterns the head word is a verb, and the entities are either subjects, or objects, or connected to the verb or to its object via a sequence of preposition phrases.
- For example:
- X/Org ← *s* — acquired — *c* → Y/Org
- X/Org ← *s* — merged ← *m* — with — *c* → Y/Org
- X/Org ← *s* — completed — *c* → acquisition ← *m* — of — *c* → Y/Org
- The link types:
- *s* - subject
- *c* - complement
- *m* - modifier

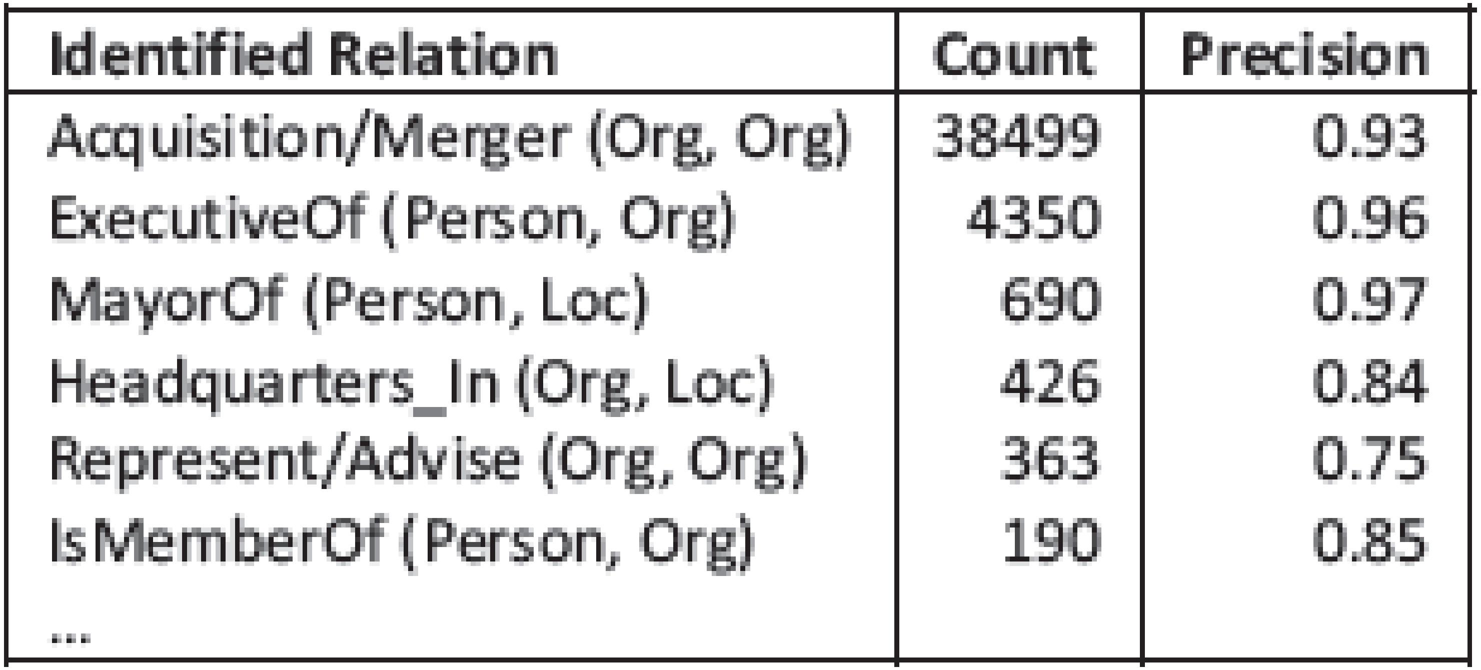
Noun based patterns

- The noun-based patterns are headed by a noun, and the entities are connected via preposition phrases, possessives, or compounds.
- For example:
- (acquisition $\leftarrow$ *m* — of — *c* $\rightarrow$ X/Org) $\leftarrow$ *m* — by — *c* $\rightarrow$ Y/Org
- X/Org $\leftarrow$ *poss* — 's — *m* $\rightarrow$ acquisition $\leftarrow$ *m* — by — *c* $\rightarrow$ Y/Org
- Merger $\leftarrow$ *m* — of — *c* $\rightarrow$ (X/Org $\leftarrow$ *conj* — and — *conj* $\rightarrow$ Y/Org).

BE patterns

- BE-patterns are headed by the verb “be” in its predicative (non-auxiliary) sense:
- X/Person (is) ← mayor ← *m* — of — *c* → Y/Loc
- X/Person , ← mayor ← *m* — of — *c* → Y/Loc
- X/Org ← *poss* — ’s — *m* → headquarters (are) ← in — *c* → Y/Loc

Experimental evaluation – ACMM corpus



Identified Relation	Count	Precision
Acquisition/Merger (Org, Org)	38499	0.93
ExecutiveOf (Person, Org)	4350	0.96
MayorOf (Person, Loc)	690	0.97
Headquarters_In (Org, Loc)	426	0.84
Represent/Advise (Org, Org)	363	0.75
IsMemberOf (Person, Org)	190	0.85
...		

Comparison to text runner

- The TextRunner query for “is mayor” produces such results as:

1. “LDA” from the sentence “*The LDA is the Mayor’s agency for business and jobs.*”
2. “year” from “*Later this year he is elected Mayor for the first time.*”
3. “bright spots” from “*Look around, you'll see bright spots in city Morganton News Herald - The following is Morganton Mayor Mel Cohen's state of the city speech.*”

- Even when the query is constrained to “Person is mayor of City”, which reduces the recall from several thousands to just 33 instances, the last sentence from the three above still remains.

Visual CARE

The *silver bullet* for the content industry

- unsupervised machine learning
- deep linguistic analysis
- automatic pattern discovery
- automatic rule definition

Accuracy above 95% vs. the industry standard of 80%

Visual CARE – the rules are written for you!

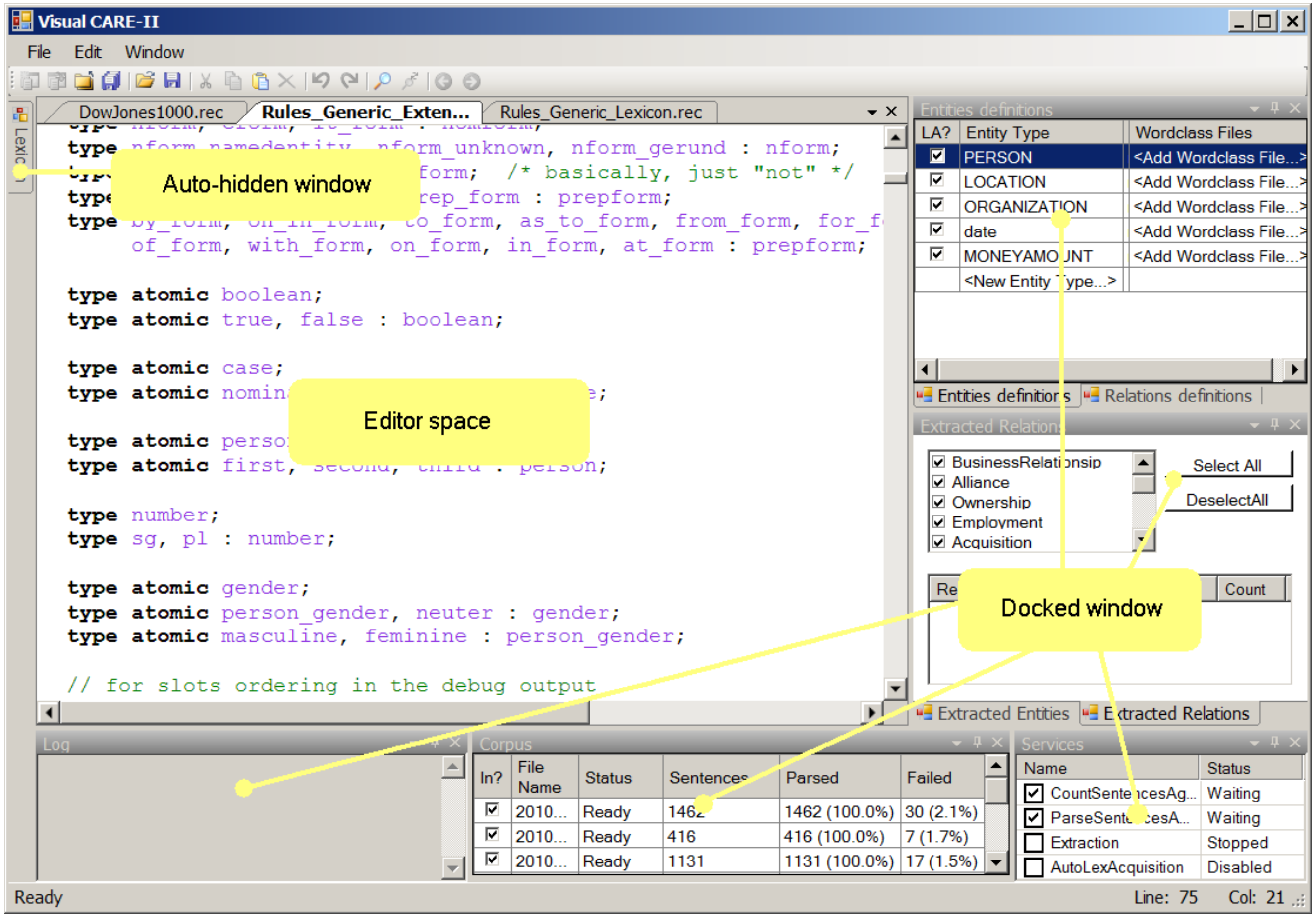
- in an intuitive language
- review and identify the rules relevant to you
- keep and refine those you care about and discard the rest
- merge similar rules to a single concept

Visual CARE – game changing benefits

- development time reduced by 99%
- discover the rules you didn’t know to look for!
- unprecedented accuracy of over 95%
- domain agnostic

A disruptive change in the way industries are able to utilize content

Visual Care Main window



```
type nform_namedentity, nform_unknown, nform_gerund : nform;
                       form;  /* basically, just "not" */
type                   rep_form : prepform;
type by_form, on_in_form, to_form, as_to_form, from_form, for_f
     of_form, with_form, on_form, in_form, at_form : prepform;

type atomic boolean;
type atomic true, false : boolean;

type atomic case;
type atomic nomin                                    e;

type atomic perso
type atomic first, second, third : person;

type number;
type sg, pl : number;

type atomic gender;
type atomic person_gender, neuter : gender;
type atomic masculine, feminine : person_gender;

// for slots ordering in the debug output
```

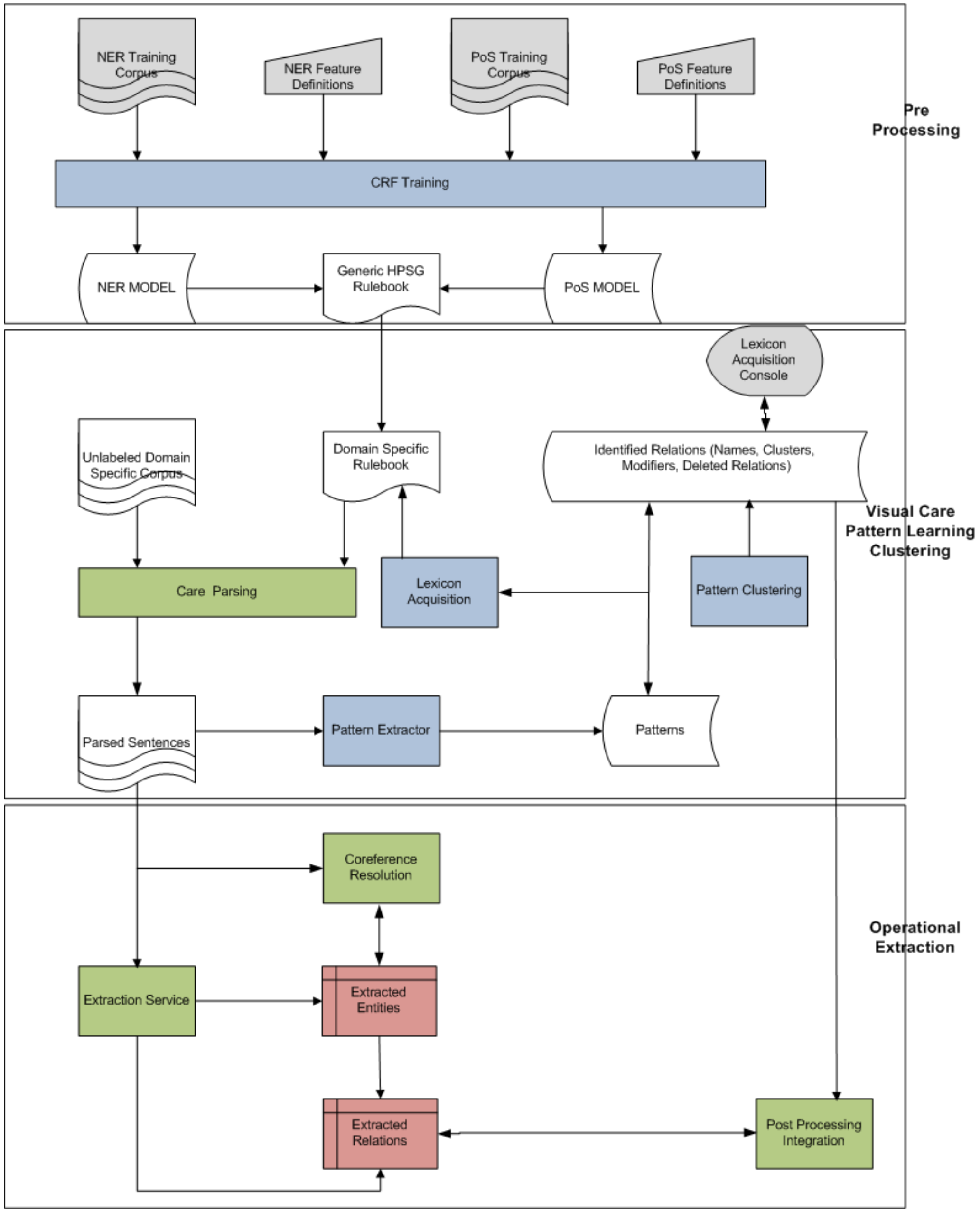
LA?	Entity Type	Wordclass Files
☑	PERSON	<Add Wordclass File...>
☑	LOCATION	<Add Wordclass File...>
☑	ORGANIZATION	<Add Wordclass File...>
☑	date	<Add Wordclass File...>
☑	MONEYAMOUNT	<Add Wordclass File...>
	<New Entity Type...>	

- ☑ BusinessRelationship
- ☑ Alliance
- ☑ Ownership
- ☑ Employment
- ☑ Acquisition

In?	File Name	Status	Sentences	Parsed	Failed
☑	2010...	Ready	1462	1462 (100.0%)	30 (2.1%)
☑	2010...	Ready	416	416 (100.0%)	7 (1.7%)
☑	2010...	Ready	1131	1131 (100.0%)	17 (1.5%)

Name	Status
☑ CountSentencesAg...	Waiting
☑ ParseSentencesA...	Waiting
☐ Extraction	Stopped
☐ AutoLexAcquisition	Disabled

Visual Care Architecture



Extractions: Medical forums

- *I was on adderall which was great, but would give me a stomach ache for a short time after each dose, and bad night sweats*
- SideEffect:
- Drug = "*adderall*"
- Symptom = ["*stomach ache*",
- "*bad night sweats*"]
- *the dr wants me to go off the avandamet and just take straight metformin for a week to see if it still causes me nausea*
- DrugReplacement:
- Drug = "*avandamet*"
- Replacement = "*metformin*"
- Reason="*to see if it still causes me nausea*"

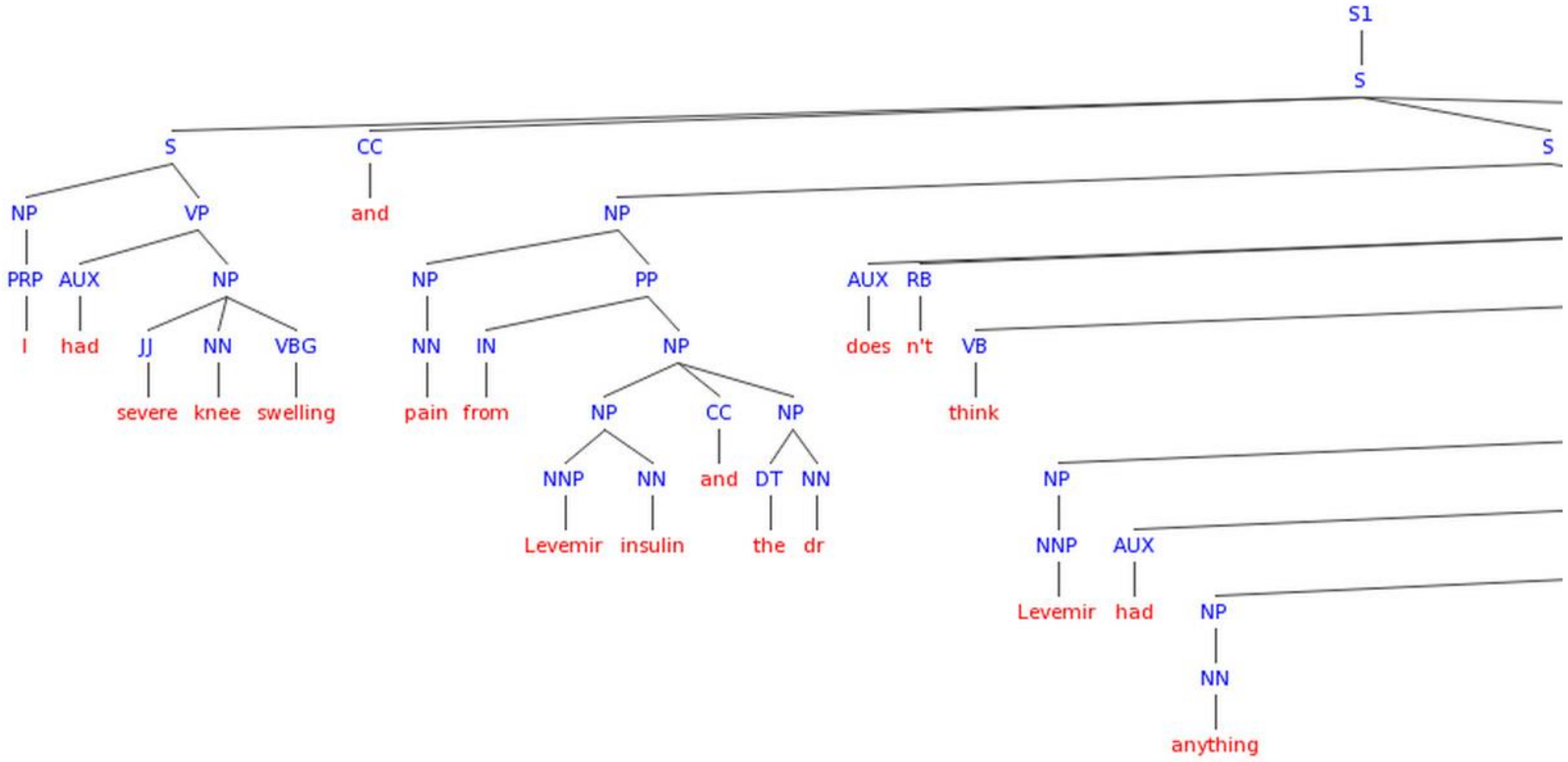
Hpsg parser

- CARE-II-HPSG is an English grammar written for the CARE-II framework, based on the principles of HPSG grammar theory.
- The grammar's lexicon is largely underspecified. Only the most frequent and functional words have full definitions, while the open classes of words are defined using generic underspecified lexical entries and tightly-integrated feature-rich sequence classification models for part-of-speech tagging (PoS) and named entity recognition (NER)
- For any input sentence, the parser generates a single highest-weight parse – the parse which is the most consistent with both the grammar rules and the NER and PoS classifiers.

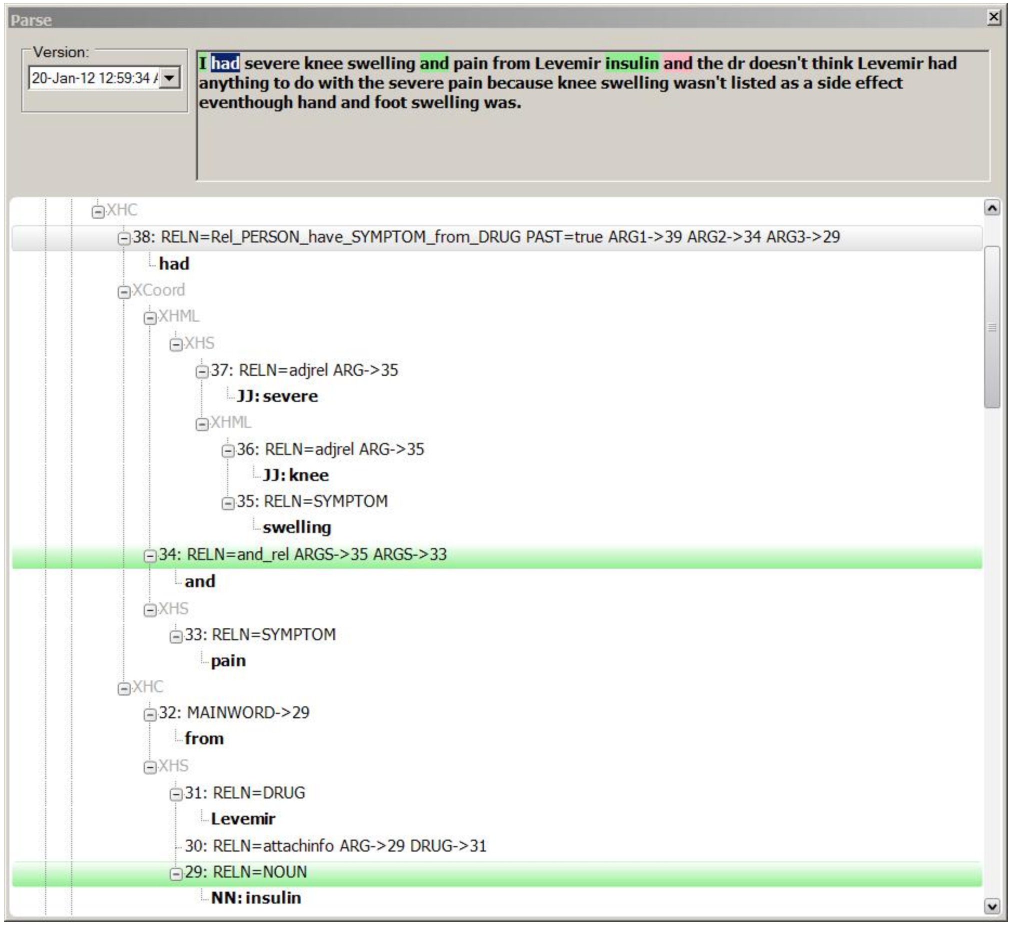
sentences I Problematic

- *I had severe knee swelling and pain from Levemir insulin and the dr doesn't think Levemir had anything to do with the severe pain because knee swelling wasn't listed as a side effect even though hand and foot swelling was.*
- The Charniak's parser missed the important domain-specific relation between "*knee swelling*" and "*Levemir insulin*" by forming a noun phrase "*pain from Levemir insulin and the dr*" and interpreting it as the subject of "*doesn't think*".

Charniak’s Parse



VC's Parse



20-Jan-12 12:59:34 /

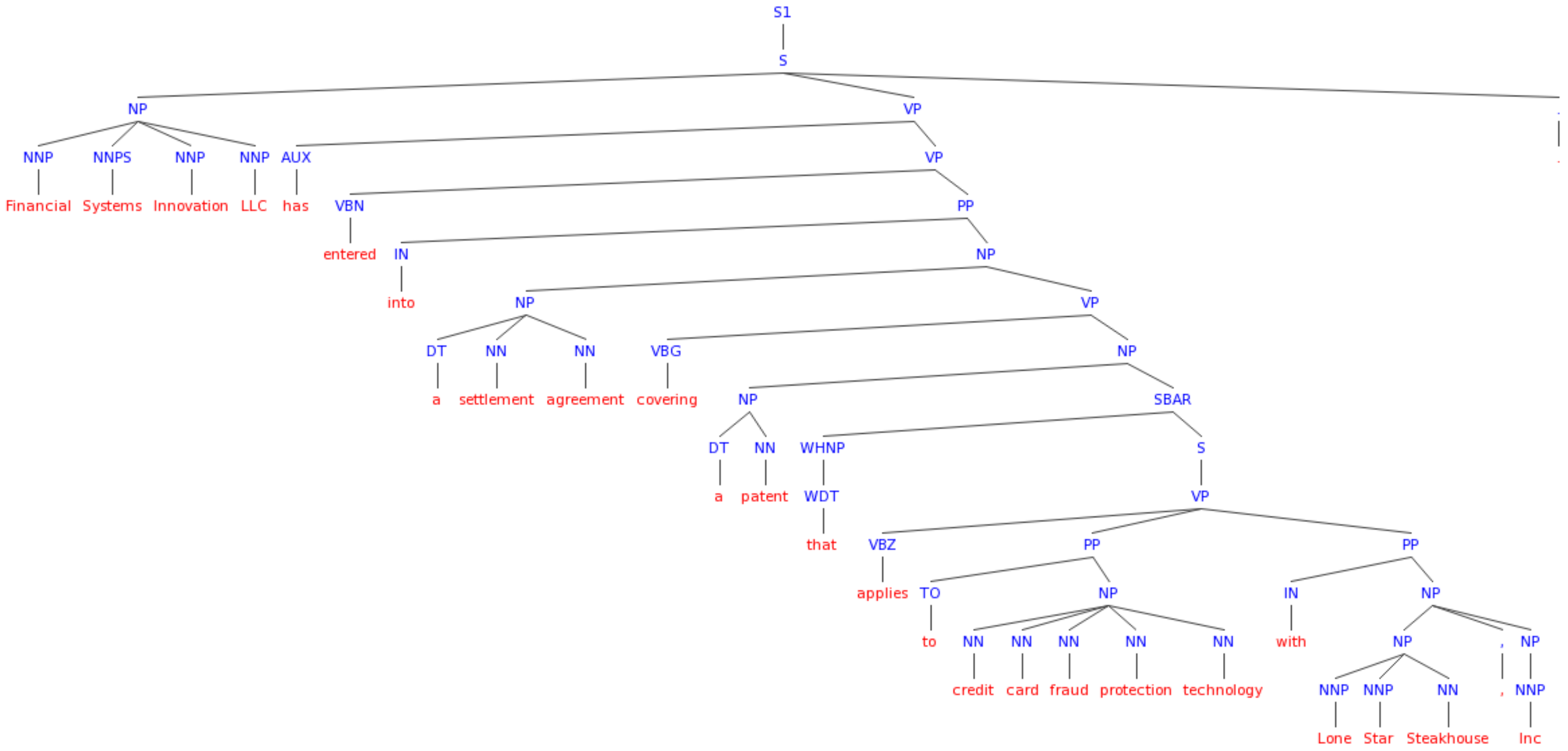
I had severe knee swelling and pain from Levemir insulin and the dr doesn't think Levemir had anything to do with the severe pain because knee swelling wasn't listed as a side effect eventhough hand and foot swelling was.

```
XHC
  38: RELN=Rel_PERSON_have_SYMPTOM_from_DRUG PAST=true ARG1->39 ARG2->34 ARG3->29
    had
  XCoord
    XHML
      XHS
        37: RELN=adjrel ARG->35
          JJ: severe
        XHML
          36: RELN=adjrel ARG->35
            JJ: knee
          35: RELN=SYMPTOM
            swelling
    34: RELN=and_rel ARGS->35 ARGS->33
      and
    XHS
      33: RELN=SYMPTOM
        pain
  XHC
    32: MAINWORD->29
      from
    XHS
      31: RELN=DRUG
        Levemir
      30: RELN=attachinfo ARG->29 DRUG->31
      29: RELN=NOUN
        NN: insulin
```

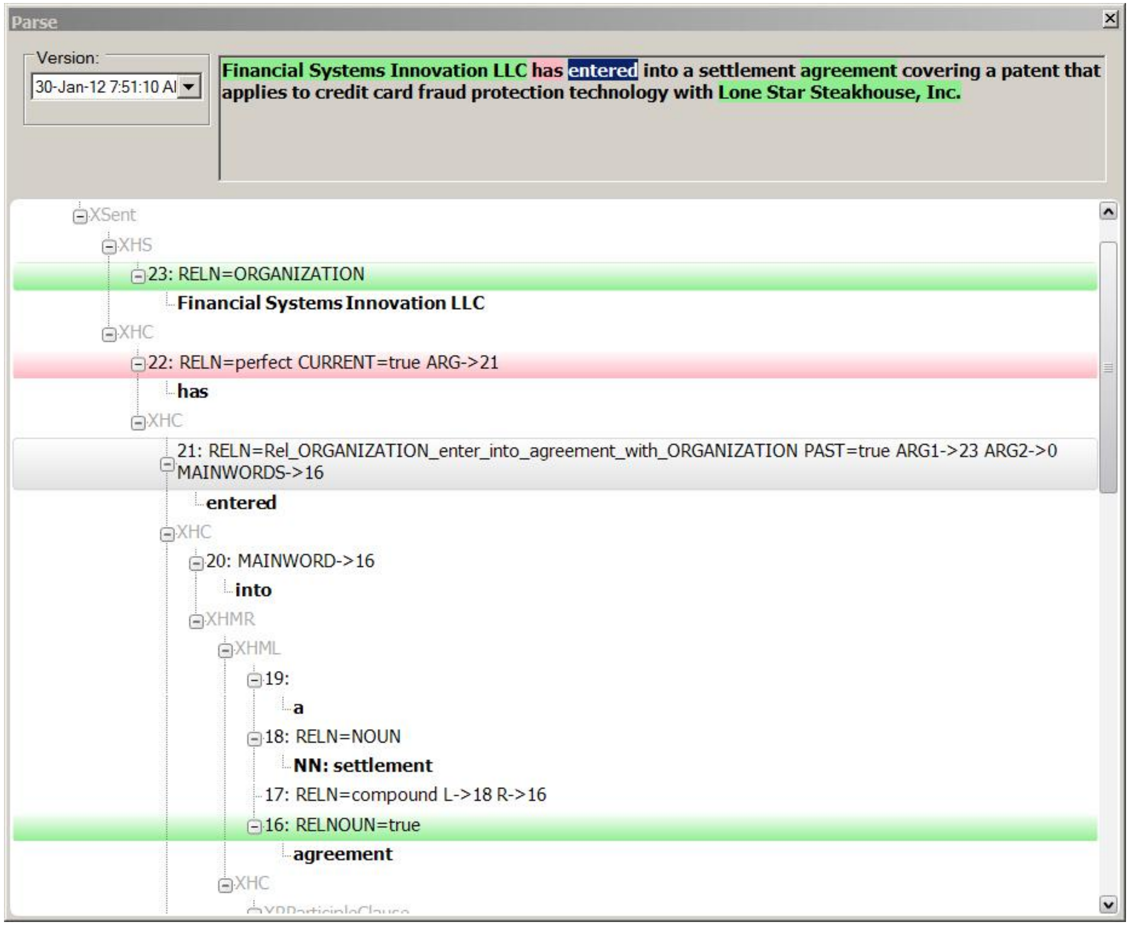
sentences II Problematic

- *Financial Systems Innovation LLC has entered into a settlement agreement covering a patent that applies to credit card fraud protection technology with Lone Star Steakhouse, Inc.*
- The wide-coverage parser makes a PP-attachment mistake, attaching "*with Lone Star Steakhouse, Inc*" to the immediately preceding NP instead of "*settlement agreement*".

Charniak’s Parse



VC’s parse



Version: 30-Jan-12 7:51:10 AI

Financial Systems Innovation LLC has entered into a settlement agreement covering a patent that applies to credit card fraud protection technology with Lone Star Steakhouse, Inc.

```
XSent
  XHS
    23: RELN=ORGANIZATION
      Financial Systems Innovation LLC
  XHC
    22: RELN=perfect CURRENT=true ARG->21
      has
    XHC
      21: RELN=Rel_ORGANIZATION_enter_into_agreement_with_ORGANIZATION PAST=true ARG1->23 ARG2->0
      MAINWORDS->16
        entered
      XHC
        20: MAINWORD->16
          into
        XHMR
          XHML
            19:
              a
            18: RELN=NOUN
              NN: settlement
            17: RELN=compound L->18 R->16
            16: RELNOUN=true
              agreement
          XHC
            XPParticipleClause
```

How Care handles these sentences

- The focused domain-specific lexical entries, learned automatically from simpler sentences with more straightforward parses, have higher weight than the generic lexical entries, and increase the chances of parses that contain them to win over parses containing generic words.
- In the first sentence, the important domain-specific word is the particular form of the verb "*have*", as in the pattern "*PersonX has SideEffectY*".
- In the second sentence, it is three words – " *enter*", "*agreement*", and the preposition "*with*", as in the pattern "*CompanyX enters into agreement with CompanyY*".

Generalization from simple sentences

- The entries in the domain-specific lexicon, after they were learned from simple patterns are able to perform extraction in much more general contexts.
- The words participate in all the general linguistic rules defined by the HPSG grammar, such as agreement, passive voice, negation, rearranging of preposition phrases order, conjunctions, etc.

Development cycle for a domain-specific Relation Extraction system

- A newly created VC project contains only the generic grammar and the standard set of named entities (Person, Organization, Location, and Date, available from a NER sequence classifier, CRF-trained on the data from CoNLL-2003 (Tjong, Sang et al. 2003) shared task).
- If additional domain-specific entity types are needed, their definitions must be supplied.
- A corpus of domain-related sentences must be added to the project. This starts the pattern extraction and lexicon acquisition process.
- The extracted patterns must be clustered, filtered, and optionally renamed, which completes the relation identification and lexicon acquisition processes.

Entity extraction

- There 4 different methods to define additional types of entity types:

1. Using a separately-trained NER model
2. CARE-II rules
3. CARE-II-HPSG lexicon definitions
4. Lists of allowed values (for entity types for which the sets of entities are closed).

- Arbitrary mixing of these methods is also possible and effective.

- drug names NER in the medical domain

- For the purposes of extracting *DrugReplacement*-s and *DrugSideEffect*-s the following phrases are all equivalent: "*Byetta*", "*Byetta pill*", "*a dose of about 100 mg of Byetta a day*", etc.
- However, in terms of the generic CARE-II-HPSG grammar, they are not equivalent, because the heads of the noun phrases are different: "*Byetta*" (DRUG), "*pill*" (generic common noun), and "*dose*" (different generic common noun), respectively.
- In order to make the longer phrases equivalent to a simple DRUG entity, it is sufficient to add the lexical entries for the possible head words: "*pill*", "*dose*", "*mg*", and several others.

NER in the medical domain - symptoms

- Extracting SYMPTOM-s is a more complex process. As with the drugs, we started by building a list of known symptoms, downloading it from www.drugs.com. Each drug there has a description page with list of possible symptoms for it.
- We combined the lists from all of the listed drugs. The resulting dictionary was used to create a set of rules in the CARE-II framework. These rules break down the symptoms to their components:

1. the *problem nouns* (e.g., "*acid*", "*bleeding*", "*ache*")
2. *problem adjectives* ("*abnormal*", "*allergic*")
3. *body_part* ("*abdomen*", "*ankle*")
4. *behavior* ("*appetite*", "*balance*", "*mood*", "*sleep*").

- These components were then added as domain-specific lexical entries, with the semantics that would allow them to form full symptom names by combining with each other in any syntactically-licensed manner.

Handling negation

- *Nestle S.A. (NESN.VX), the world's largest food and beverages producer, Tuesday said it won't bid for U.K.-based confectionery company Cadbury Plc (CBY).*
- The syntax and semantics of such forms of negation are handled in the generic grammar in a way compatible with the HPSG grammar theory.
- In practice, if either the main verb or one of the slots of a relation is modified by a negating modifier ("*not*" and its various forms, and other negating words), then the extracted relation is marked as negated.

Domain specific lexicon acquisition

- We add to the project an unlabeled corpus – a set of domain-relevant sentences – and run the pattern learning and lexicon acquisition process. VC uses the unlabeled corpus for discovering linguistic patterns that can be directly translated into CARE-II-HPSG lexicon definitions.
- The patterns whose corpus frequency exceeds the threshold (which is set to 2 in this study) get "promoted" into lexicon definitions, added to the project, and the affected parts of the corpus are automatically reparsed.

example

- *The Rel_ORG_enter_into_agreement_with_ORG* pattern adds four entries: the prepositions "*into*" and "*with*", the noun "*agreement*", and the verb "*enter*".

1. "*Into*" and "with" are defined as argument prepositions
2. "*agreement*" is defined as a noun with a special SYN.HEAD.FORM and without any non-generic semantics
3. "*enter*" is defined as a verb with three complements and with the output relation semantics.

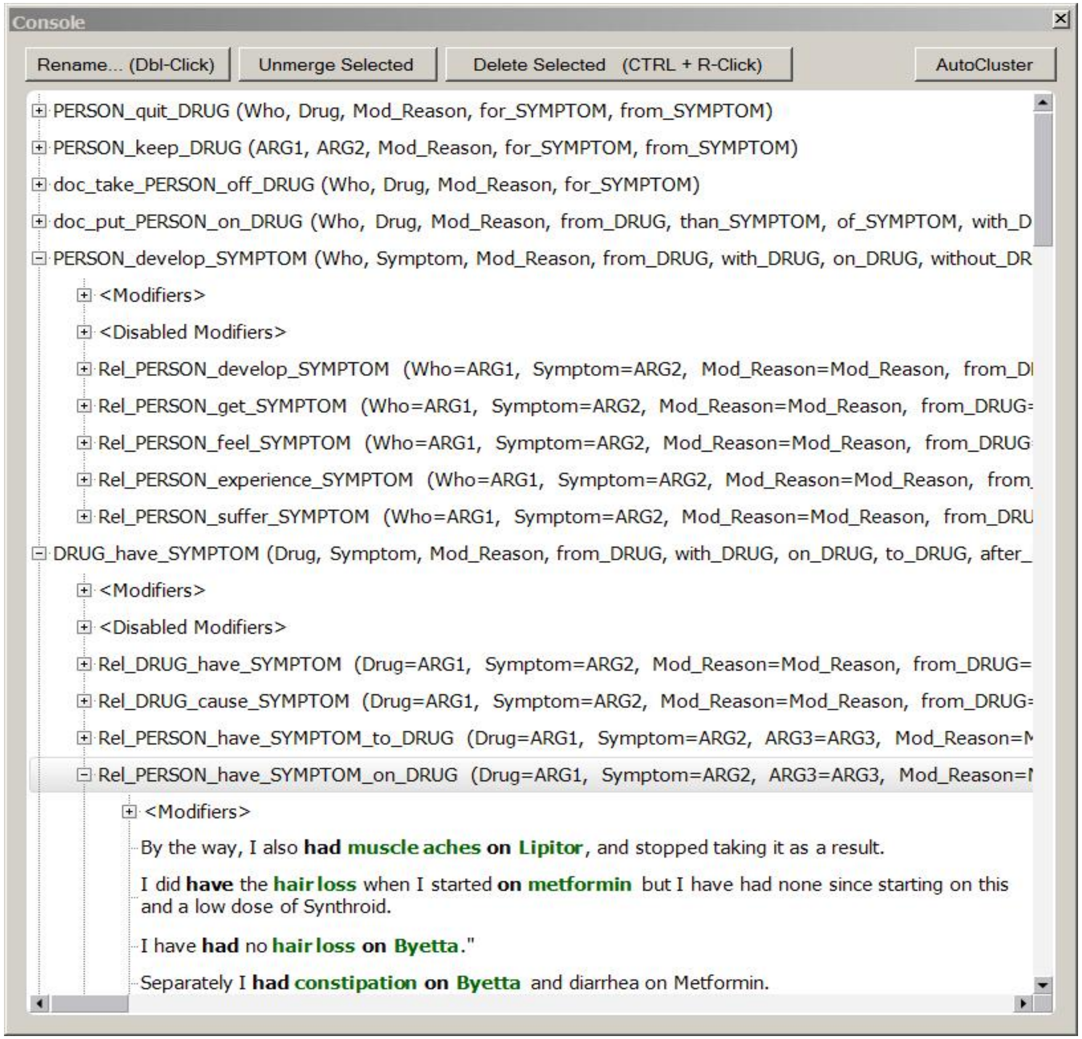
- The definitions of non-pattern-specific words, such as "*into*", "*with*" and "*agreement*" can be reused by many patterns.
- The pattern names show only the complements – the required pieces of the pattern. The actual instances of the relation may also contain optional pieces, specified by modifiers.

Relation clustering

- The automatic pattern clustering uses a variant of the HAC (Hierarchical Agglomerative Clustering) algorithm with single linkage, which was shown in (Rosenfeld and Feldman 2007) to be superior for the relation identification task.
- The direct similarity estimation between patterns takes into account:

1. the structural similarity between patterns, including standard syntactic transformations;
2. identity of slots' entity types;
3. identity or synonymy or other association between specific words in the patterns, discovered using WordNet (Fellbaum 1998).

Clustering example



```
⊞ PERSON_quit_DRUG (Who, Drug, Mod_Reason, for_SYMPTOM, from_SYMPTOM)
⊞ PERSON_keep_DRUG (ARG1, ARG2, Mod_Reason, for_SYMPTOM, from_SYMPTOM)
⊞ doc_take_PERSON_off_DRUG (Who, Drug, Mod_Reason, for_SYMPTOM)
⊞ doc_put_PERSON_on_DRUG (Who, Drug, Mod_Reason, from_DRUG, than_SYMPTOM, of_SYMPTOM, with_D
⊟ PERSON_develop_SYMPTOM (Who, Symptom, Mod_Reason, from_DRUG, with_DRUG, on_DRUG, without_DR
    ⊞ <Modifiers>
    ⊞ <Disabled Modifiers>
    ⊞ Rel_PERSON_develop_SYMPTOM (Who=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from_DI
    ⊞ Rel_PERSON_get_SYMPTOM (Who=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from_DRUG=
    ⊞ Rel_PERSON_feel_SYMPTOM (Who=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from_DRUG
    ⊞ Rel_PERSON_experience_SYMPTOM (Who=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from
    ⊞ Rel_PERSON_suffer_SYMPTOM (Who=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from_DRU
⊟ DRUG_have_SYMPTOM (Drug, Symptom, Mod_Reason, from_DRUG, with_DRUG, on_DRUG, to_DRUG, after_
    ⊞ <Modifiers>
    ⊞ <Disabled Modifiers>
    ⊞ Rel_DRUG_have_SYMPTOM (Drug=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from_DRUG=
    ⊞ Rel_DRUG_cause_SYMPTOM (Drug=ARG1, Symptom=ARG2, Mod_Reason=Mod_Reason, from_DRUG=
    ⊞ Rel_PERSON_have_SYMPTOM_to_DRUG (Drug=ARG1, Symptom=ARG2, ARG3=ARG3, Mod_Reason=M
    ⊟ Rel_PERSON_have_SYMPTOM_on_DRUG (Drug=ARG1, Symptom=ARG2, ARG3=ARG3, Mod_Reason=I
        ⊞ <Modifiers>
        By the way, I also had muscle aches on Lipitor, and stopped taking it as a result.
        I did have the hair loss when I started on metformin but I have had none since starting on this
        and a low dose of Synthroid.
        I have had no hair loss on Byetta."
        Separately I had constipation on Byetta and diarrhea on Metformin.
```

Extending relations by utilizing modifiers

- The lexicon acquisition component is currently able to identify and learn two common modifier patterns: preposition phrase modifiers and possessive construction modifiers.
- Preposition phrase (PP) modifiers have the form of a preposition complemented by a noun phrase (<prep> NP), and can modify verb and noun phrases. The generic CARE-II-HPSG grammar contains default lexical entries for all prepositions, which allow any PP to modify any NP or VP.
- Possessive construction modifiers always modify noun phrases. They have three syntactically different but semantically identical forms: possessive determiner form (X's <noun>), compound noun form (X <noun>), and *of*-preposition phrase form (*of* X).

examples

- In order for a non-default domain-specific modifier to be useful, and therefore learnable, its NP part must contain a relation argument – extractable entity – either directly, as in:
- *In January 1997, Hays bought German distributor Daufenbach for 30 million GBP, ...*
- or via a PP-NP chain, as in:
- *Ms. Bennett succeeds William J. Viveen, Jr., as a member of the Interleukin Board of Directors.*
- assuming the verb relation patterns *Rel_ORG_buy_ORG* and *Rel_PERSON_succeed_PERSON* are already learned, the two sentences above would generate the PP modifier patterns *Mod_verb_in_DATE* and *Mod_verb_as_member_of_ORG*.

Modifiers are cross-pattern

- The same PP modifier can attach to phrases extracting different relations, and the same possessive modifier can attach to nouns of the same type within different patterns.
- It is possible to precisely fine-tune the scope of a modifier to specify the relations and clusters to which the modifier is applicable, and, for each relation and cluster, the slot name of the argument that the modifier extracts.
- Thus, the scope of *Mod_verb_as_member_of_ORG* can be limited to the *ManagementChange* cluster, and the extracted slot can be specified as EMPLOYER.

Co-reference resolution

- we implemented the co-reference resolution system of (Lee, Peirsman et al. 2011), which was the best-performing in CoNLL-2011 shared task, and which is perfectly adaptable to the CARE-II-HPSG/VC environment.
- The method for resolution of co-references is based on locating all noun phrases, identifying their properties, and then clustering them in several deterministic iterations (called *sieves*), starting with the highest-confidence rules and moving to lower-confidence higher-recall ones. In each iteration, the order of candidate checks is deterministic, and any matching noun phrases with matching properties are immediately clustered together.
- Despite its simplicity, the method showed state-of-the-art accuracy, outperforming the other systems in the CoNLL shared task.

Co-reference resolution

- The method is especially suitable for the our framework, because all information the method requires is already extracted: the noun phrases are located from the parses, together with their properties, which are identified from HPSG feature structures.
- Co-reference resolution module tries to resolve all noun phrases, although non-entity ones are discarded. This is necessary for improving the accuracy on the relevant entity mentions, by removing irrelevant candidates.

Examples

- The entity to resolve to is the first found one with matching features. The order of checking is strictly by distance (closest first), except for pronominal/generic-nominal references, for which
  the entities in subject positions are checked before the others.

  Microsoft announced it plans to acquire Visio. The company said it will finalize its plans within a week.

  Only pronomial and generic nominal references are here. The first "it“ straightforwardly matches "Microsoft". "The company" also matches "Microsoft" first, because it's a pronominal reference, and "Microsoft" is in the subject position. The second "it" matches "the company", and so is also resolved to "Microsoft".

  Mark said that he used Symlin and it caused him to get a rash. He said that it bothered him.

  First "he" resolves to "Mark". "It" does not match "Mark", and so is resolved to "Symlin". Other "Him" and "he" are also resolved to "Mark". The second "it" is resolved to "a rash", because it is closer and neither of the matching candidates - "Symlin" and "a rash" - is in the subject position.

Post processing

- The post-processing phase tried to combines a few simple relation to one complex relation.
- The slots of any adjacent relation-pieces are merged, if the pieces are compatible, and if none of the slots contradict each other.
- For the *DrugReplacement* relation, *StopUsingDrug* can merge with *StartUsingDrug* and with *UseDrug*. For the *SideEffect* relation, *HasSymptom* can merge with any of the three: *StartUsingDrug*, *StopUsingDrug*, or *UseDrug*.
- After merging, all *StartUsingDrug*, *UseDrug*, and *HasSymptom* relations that were not identified as parts of a larger relation can be removed, since we are interested only in the full *DrugReplacement* and *SideEffect* relations in this project.

evaluation

- The input corpus for the medical domain consists of about 650,000 messages downloaded from the archives of several public medical forums.
- The contents of the messages were separated into sentences, filtered and deduplicated, leaving about 300,000 sentences totaling in about 40MB of ASCII text.

Evaluation results

	Without Coreference			With Coreference		
Relation	Precision	Count	Recall est.	Prec	Count	Recall est.
DrugReplacement	0.89	2732	0.65	0.88	3039	0.74
Slot: Drug	0.96	2850	0.98	0.95	3163	0.98
Slot: Replacement	0.93	638	0.70	0.93	825	0.80
Slot: Reason	0.93	294	0.76			
SideEffect	0.85	2278	0.66			
Acquisition	0.88	191	0.48	0.85	341	0.81

Evaluation methodology

- Precision
  - **We evaluated on a random set of 200 extractions.**
- Recall
  - **We checked only the sentences that contained at least two DRUG entities for the *DrugReplacement* and at least one DRUG and at least one SYMPTOM entity for the *SideEffect*, together with an indicative word, such as "replace", "switch", "change", or one of their derivatives for the *DrugReplacement*.**

Financial domain evaluation

- In the financial domain, we used two corpora: "Corpus A" of 200,000 sentences selected randomly from Dow Jones newswire articles published between 1990 and 2010, and "Corpus B" of 40,000 sentences from Dow Jones in 2011.
- First, we used VC to learn as many relations as it could from the Corpus A. The system found 919 different patterns, of which 481, clustered into 172 relation clusters, were judged interesting (by manual inspection).
- Here are the ten top (according to the number of instances found in the corpus) identified clusters:

1. *Acquisition* (*Organization*, *Organization*)
2. *Replacement* (*Person*, *Person*)
3. *Employment* (*Person*, *Organization*)
4. *NewProduct* (*Product*, *Organization*)
5. *Agreement* (*Organization*, *Organization*)
6. *ORG_own_interest_in_ORG* (*Org*, *Org*)
7. *HeadPosition* (*Person*, *Organization*)
8. *Resignation* (*Person*, *Organization*)
9. *Sale* (*Organization*, *Organization*)
10. *HeadquartersIn*(*Organization, Location*

Coverage experiment

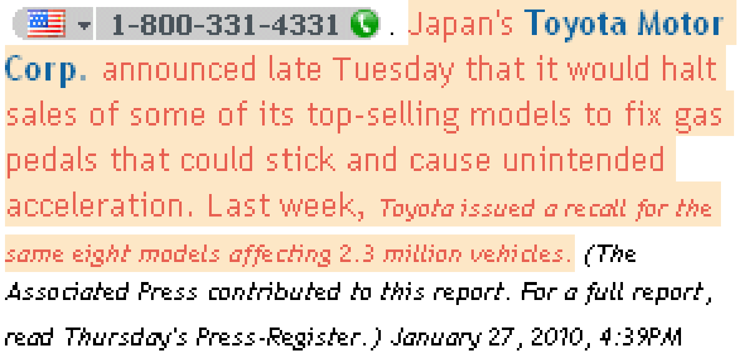
- we used VC to learn relations from Corpus B independently from Corpus A, and compared the sets of relations extracted from the Corpus B by domain-specific lexicons learned from itself and from Corpus A.
- We have found that out of 1742 relation instances found in Corpus B, we were able to extract 1367 of them (78.4%) using the patterns learned from Corpus A.

Sentiment Analysis of Stocks from News Sites

So, How Can We Utilize NLP for Making Money?

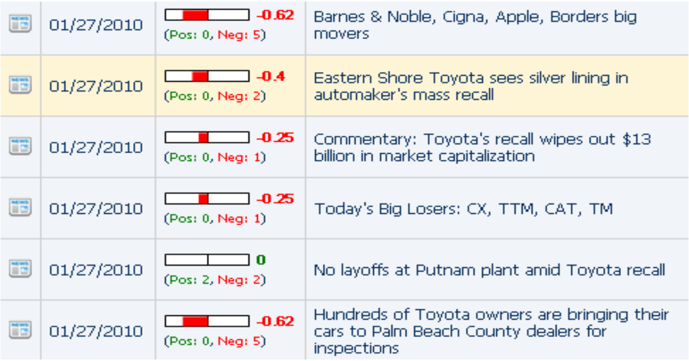
- Goal: sentiment analysis of financial texts as an aid for stock investment

1. Tagging positive and negative sentiment in articles



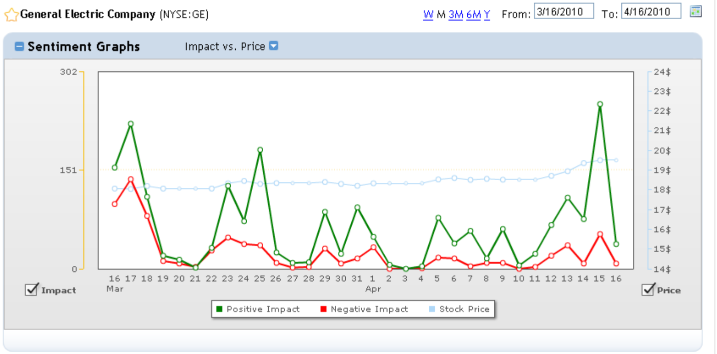
1-800-331-4331 . Japan's Toyota Motor Corp. announced late Tuesday that it would halt sales of some of its top-selling models to fix gas pedals that could stick and cause unintended acceleration. Last week, *Toyota issued a recall for the same eight models affecting 2.3 million vehicles.* *(The Associated Press contributed to this report. For a full report, read Thursday's Press-Register.) January 27, 2010, 4:39PM*

2. Article scoring



	01/27/2010	-0.62 (Pos: 0, Neg: 5)	Barnes & Noble, Cigna, Apple, Borders big movers
	01/27/2010	-0.4 (Pos: 0, Neg: 2)	Eastern Shore Toyota sees silver lining in automaker's mass recall
	01/27/2010	-0.25 (Pos: 0, Neg: 1)	Commentary: Toyota's recall wipes out $13 billion in market capitalization
	01/27/2010	-0.25 (Pos: 0, Neg: 1)	Today's Big Losers: CX, TTM, CAT, TM
	01/27/2010	0 (Pos: 2, Neg: 2)	No layoffs at Putnam plant amid Toyota recall
	01/27/2010	-0.62 (Pos: 0, Neg: 5)	Hundreds of Toyota owners are bringing their cars to Palm Beach County dealers for inspections

3. Score aggregation: daily and cumulative score



General Electric Company (NYSE:GE)

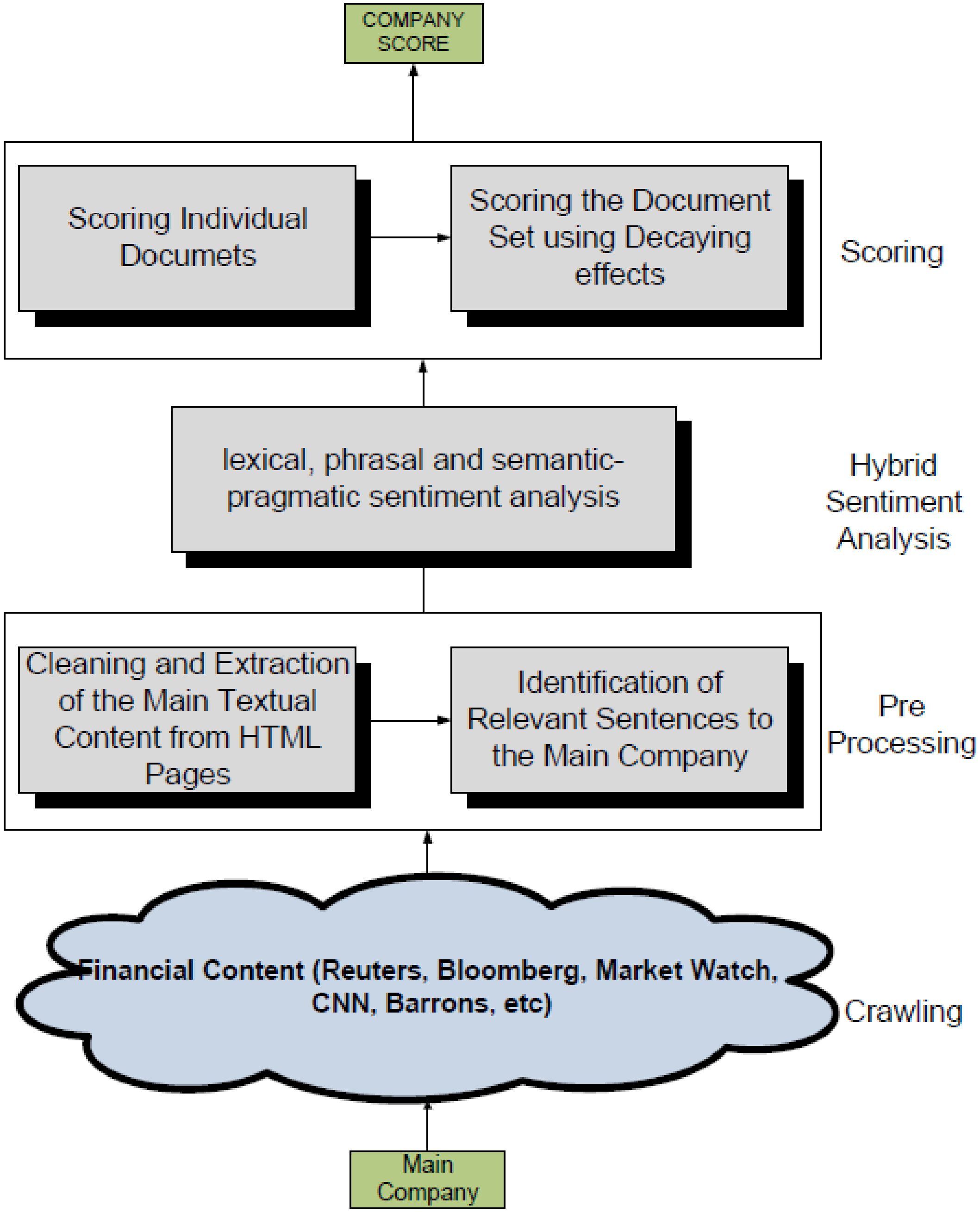
From: 3/16/2010 To: 4/16/2010

The Need for Event Based SA

Toyota announces voluntary recall of their ***highly successful top selling*** *2010 model-year cars*

- Phrase-level SA:
  - highly successful top selling ⇒ **positive**
  - Or at best neutral
    - Taking into account voluntary recall ⇒ **negative**
- Need to recognize the whole sentence as a “product recall” event!

CaRE extraction Engine



COMPANY SCORE

Scoring Individual Documets

Scoring the Document Set using Decaying effects

Scoring

lexical, phrasal and semantic-pragmatic sentiment analysis

Hybrid Sentiment Analysis

Cleaning and Extraction of the Main Textual Content from HTML Pages

Identification of Relevant Sentences to the Main Company

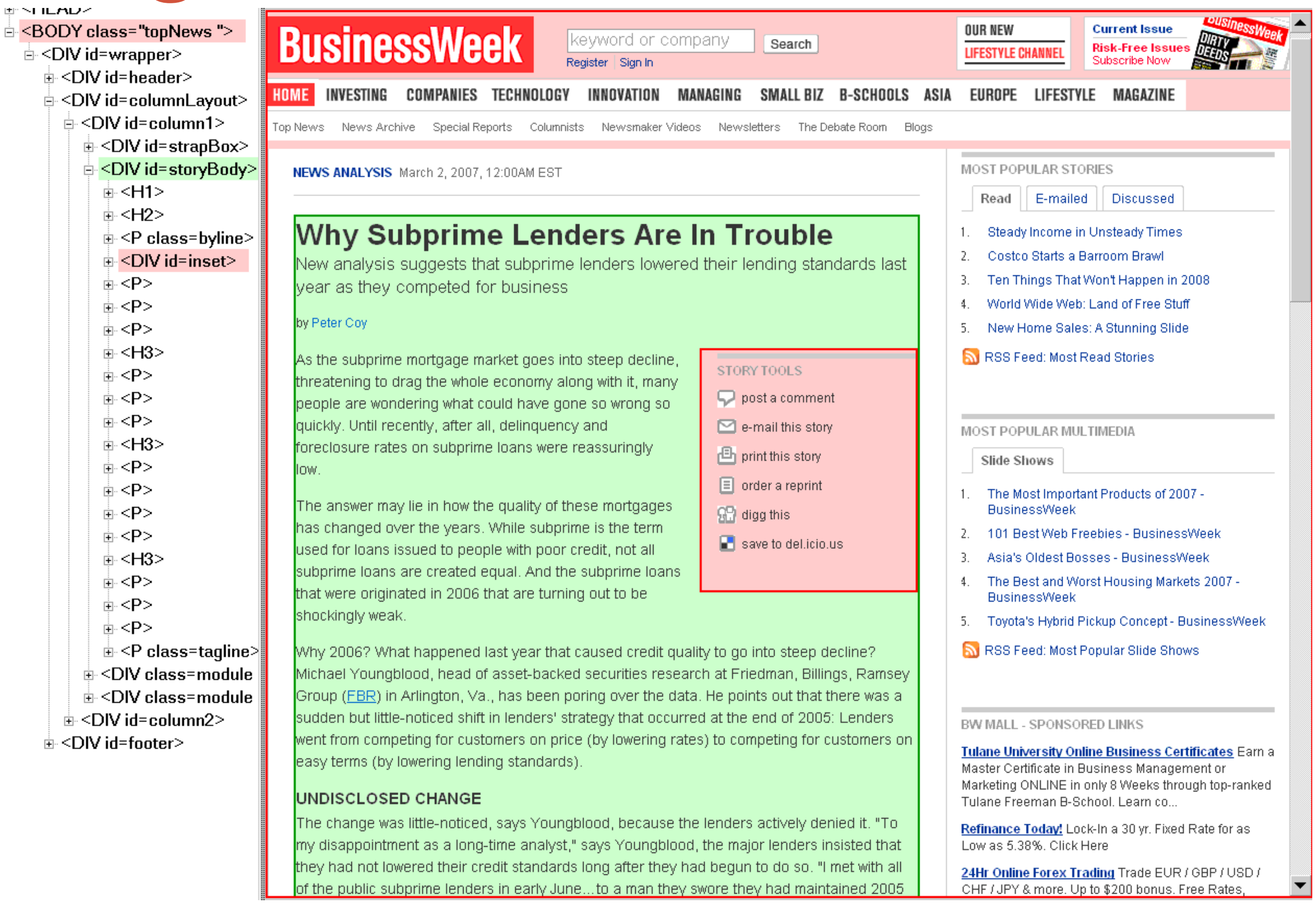
Pre Processing

Financial Content (Reuters, Bloomberg, Market Watch, CNN, Barrons, etc)

Crawling

Main Company

Template Based Approach to Content Filtering



```
<HEAD>
<BODY class="topNews ">
  <DIV id=wrapper>
    <DIV id=header>
    <DIV id=columnLayout>
      <DIV id=column1>
        <DIV id=strapBox>
        <DIV id=storyBody>
          <H1>
          <H2>
          <P class=byline>
          <DIV id=inset>
          <P>
          <P>
          <P>
          <H3>
          <P>
          <P>
          <P>
          <H3>
          <P>
          <P>
          <P>
          <P>
          <H3>
          <P>
          <P>
          <P>
          <P class=tagline>
        <DIV class=module
        <DIV class=module
      <DIV id=column2>
    <DIV id=footer>
```

NEWS ANALYSIS March 2, 2007, 12:00AM EST

Why Subprime Lenders Are In Trouble

New analysis suggests that subprime lenders lowered their lending standards last year as they competed for business

by Peter Coy

STORY TOOLS
- post a comment
- e-mail this story
- print this story
- order a reprint
- digg this
- save to del.icio.us

As the subprime mortgage market goes into steep decline, threatening to drag the whole economy along with it, many people are wondering what could have gone so wrong so quickly. Until recently, after all, delinquency and foreclosure rates on subprime loans were reassuringly low.

The answer may lie in how the quality of these mortgages has changed over the years. While subprime is the term used for loans issued to people with poor credit, not all subprime loans are created equal. And the subprime loans that were originated in 2006 that are turning out to be shockingly weak.

Why 2006? What happened last year that caused credit quality to go into steep decline? Michael Youngblood, head of asset-backed securities research at Friedman, Billings, Ramsey Group (FBR) in Arlington, Va., has been poring over the data. He points out that there was a sudden but little-noticed shift in lenders' strategy that occurred at the end of 2005: Lenders went from competing for customers on price (by lowering rates) to competing for customers on easy terms (by lowering lending standards).

UNDISCLOSED CHANGE

The change was little-noticed, says Youngblood, because the lenders actively denied it. "To my disappointment as a long-time analyst," says Youngblood, the major lenders insisted that they had not lowered their credit standards long after they had begun to do so. "I met with all of the public subprime lenders in early June...to a man they swore they had maintained 2005

MOST POPULAR STORIES

Read | E-mailed | Discussed

1. Steady Income in Unsteady Times
2. Costco Starts a Barroom Brawl
3. Ten Things That Won't Happen in 2008
4. World Wide Web: Land of Free Stuff
5. New Home Sales: A Stunning Slide

RSS Feed: Most Read Stories

MOST POPULAR MULTIMEDIA

Slide Shows

1. The Most Important Products of 2007 - BusinessWeek
2. 101 Best Web Freebies - BusinessWeek
3. Asia's Oldest Bosses - BusinessWeek
4. The Best and Worst Housing Markets 2007 - BusinessWeek
5. Toyota's Hybrid Pickup Concept - BusinessWeek

RSS Feed: Most Popular Slide Shows

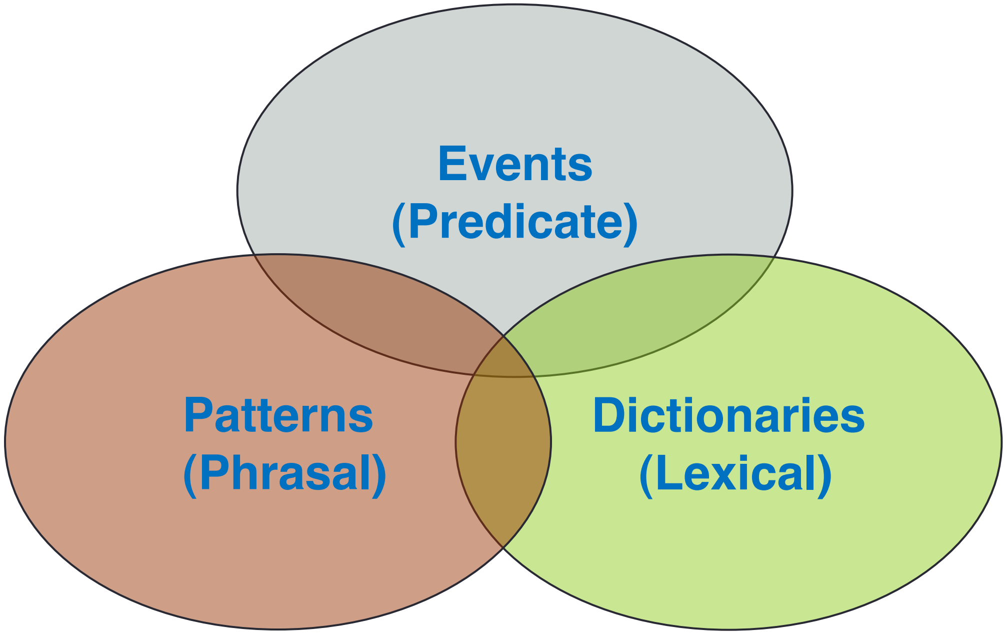
BW MALL - SPONSORED LINKS

Tulane University Online Business Certificates Earn a Master Certificate in Business Management or Marketing ONLINE in only 8 Weeks through top-ranked Tulane Freeman B-School. Learn co...

Refinance Today! Lock-In a 30 yr. Fixed Rate for as Low as 5.38%. Click Here

24Hr Online Forex Trading Trade EUR / GBP / USD / CHF / JPY & more. Up to $200 bonus. Free Rates,

Hybrid Sentiment Analysis



- All levels are part of the same rulebook, and are therefore considered simultaneously by CaRE

Dictionary-based sentiment

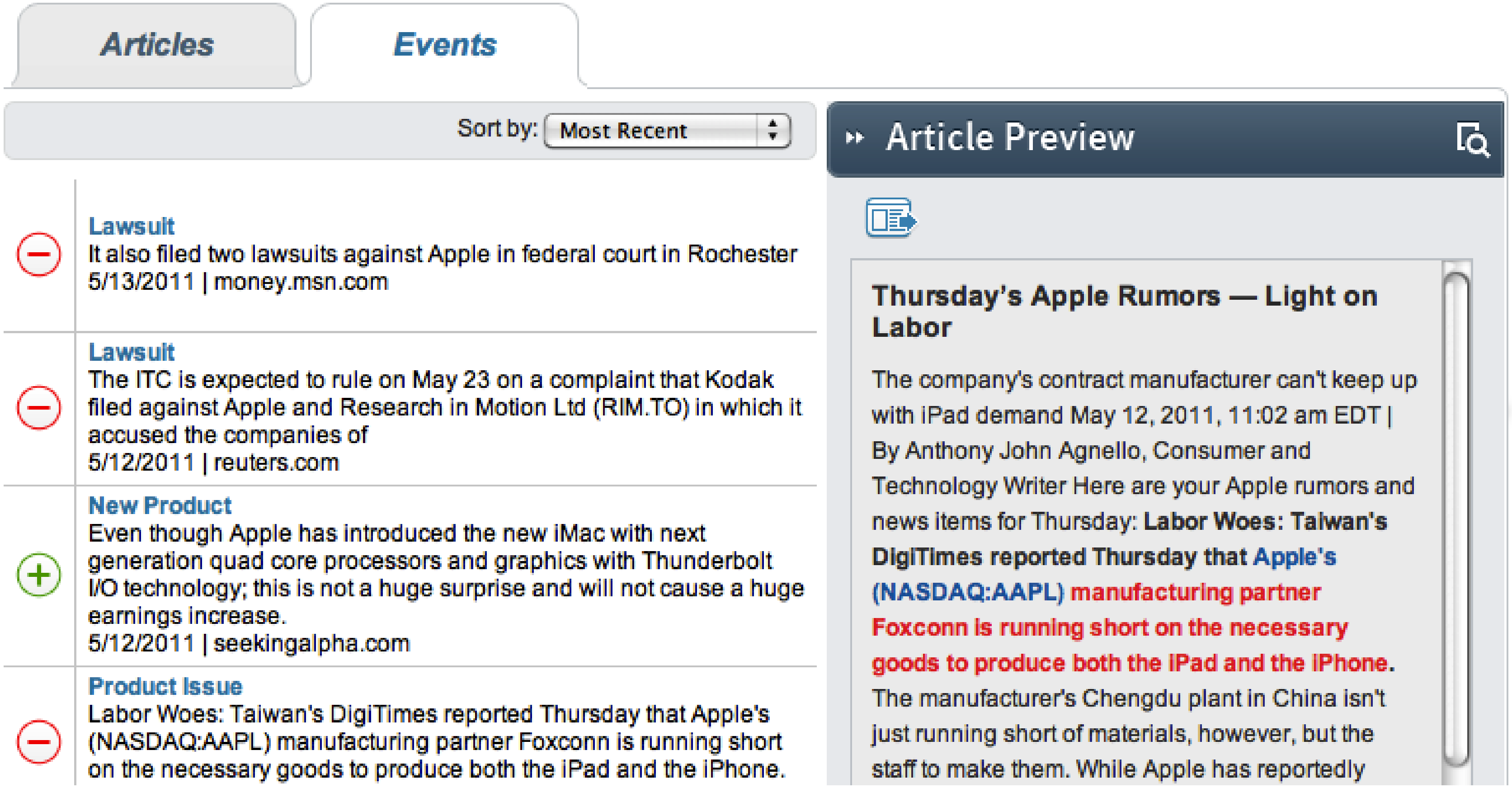
- Started with available sentiment lexicons
  - Domain-specific and general
  - Improved by our content experts
- Examples
  - Modifiers: attractive, superior, inefficient, risky
  - Verbs: invents, advancing, failed, lost
  - Nouns: opportunity, success, weakness, crisis
  - Expressions: exceeding expectations, chapter 11
- Emphasis and reversal
  - successful, **extremely successful,** far from successful

Event-Based Sentiment

- Product release/approval/recall, litigations, acquisitions, workforce change, analyst recommendations and many more
- Semantic role matters:
  - Google is being sued/is suing…
- Need to address historical/speculative events
  - Google acquired YouTube **in 2006**
  - **What if Google buys Yahoo** and the software giant Microsoft remains a single company fighting for the power of the Internet?

The Stock Sonar
See through the current

Extract sentiment and events of public companies from online news articles to enable better trading decisions



Sort by: Most Recent

⊖ **Lawsuit**
It also filed two lawsuits against Apple in federal court in Rochester
5/13/2011 | money.msn.com

⊖ **Lawsuit**
The ITC is expected to rule on May 23 on a complaint that Kodak filed against Apple and Research in Motion Ltd (RIM.TO) in which it accused the companies of
5/12/2011 | reuters.com

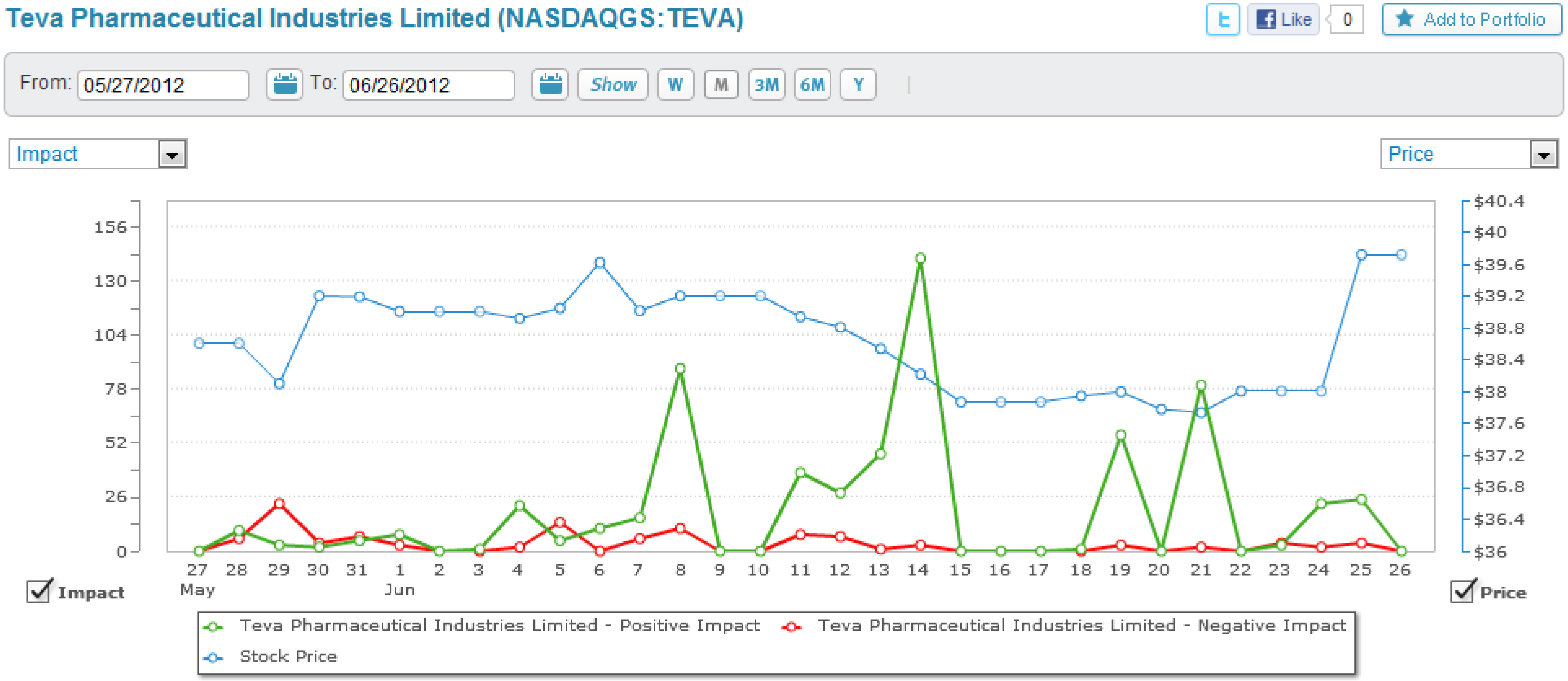
⊕ **New Product**
Even though Apple has introduced the new iMac with next generation quad core processors and graphics with Thunderbolt I/O technology; this is not a huge surprise and will not cause a huge earnings increase.
5/12/2011 | seekingalpha.com

⊖ **Product Issue**
Labor Woes: Taiwan's DigiTimes reported Thursday that Apple's (NASDAQ:AAPL) manufacturing partner Foxconn is running short on the necessary goods to produce both the iPad and the iPhone.

Thursday's Apple Rumors — Light on Labor

The company's contract manufacturer can't keep up with iPad demand May 12, 2011, 11:02 am EDT | By Anthony John Agnello, Consumer and Technology Writer Here are your Apple rumors and news items for Thursday: **Labor Woes: Taiwan's DigiTimes reported Thursday that Apple's (NASDAQ:AAPL) manufacturing partner Foxconn is running short on the necessary goods to produce both the iPad and the iPhone.** The manufacturer's Chengdu plant in China isn't just running short of materials, however, but the staff to make them. While Apple has reportedly

TEVA



Teva Pharmaceutical Industries Limited (NASDAQGS:TEVA)

From: 05/27/2012 To: 06/26/2012

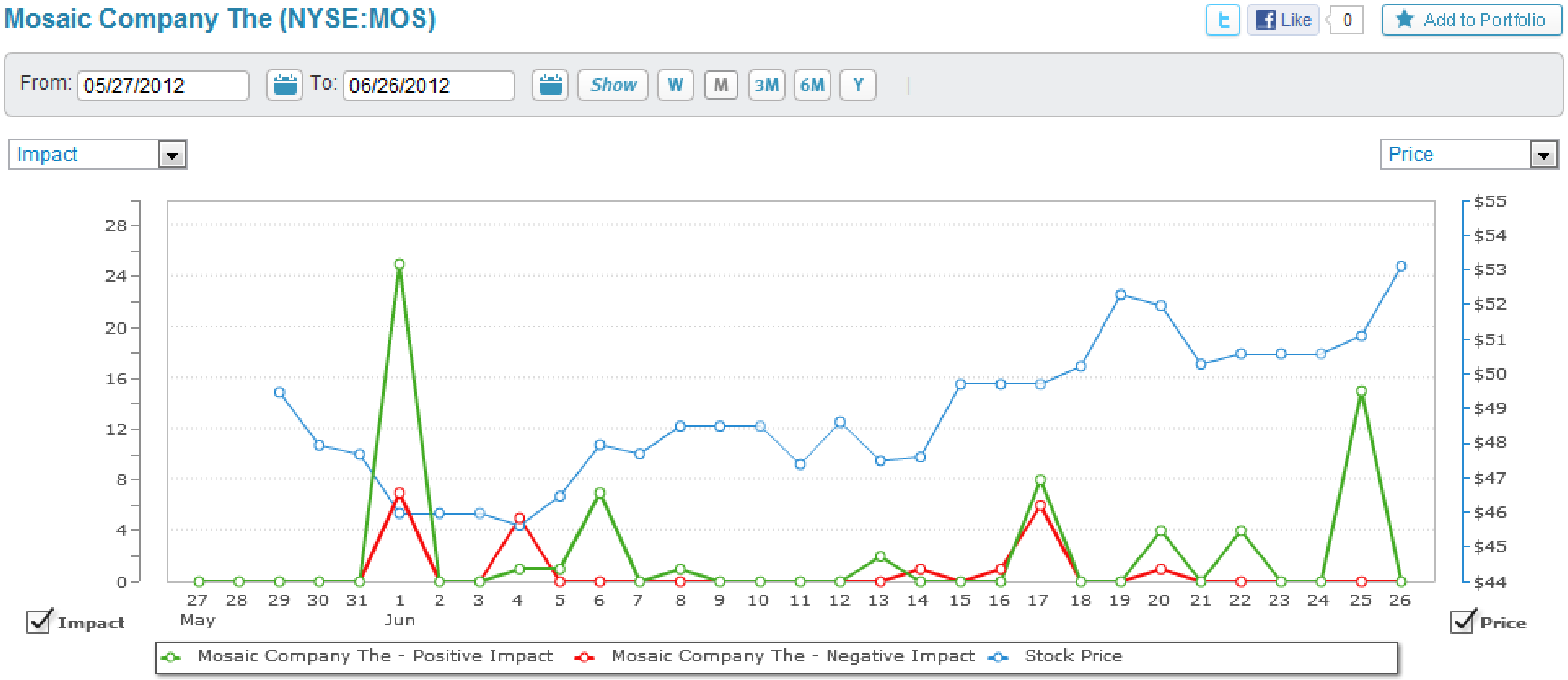
Impact

Price

☑ Impact

☑ Price

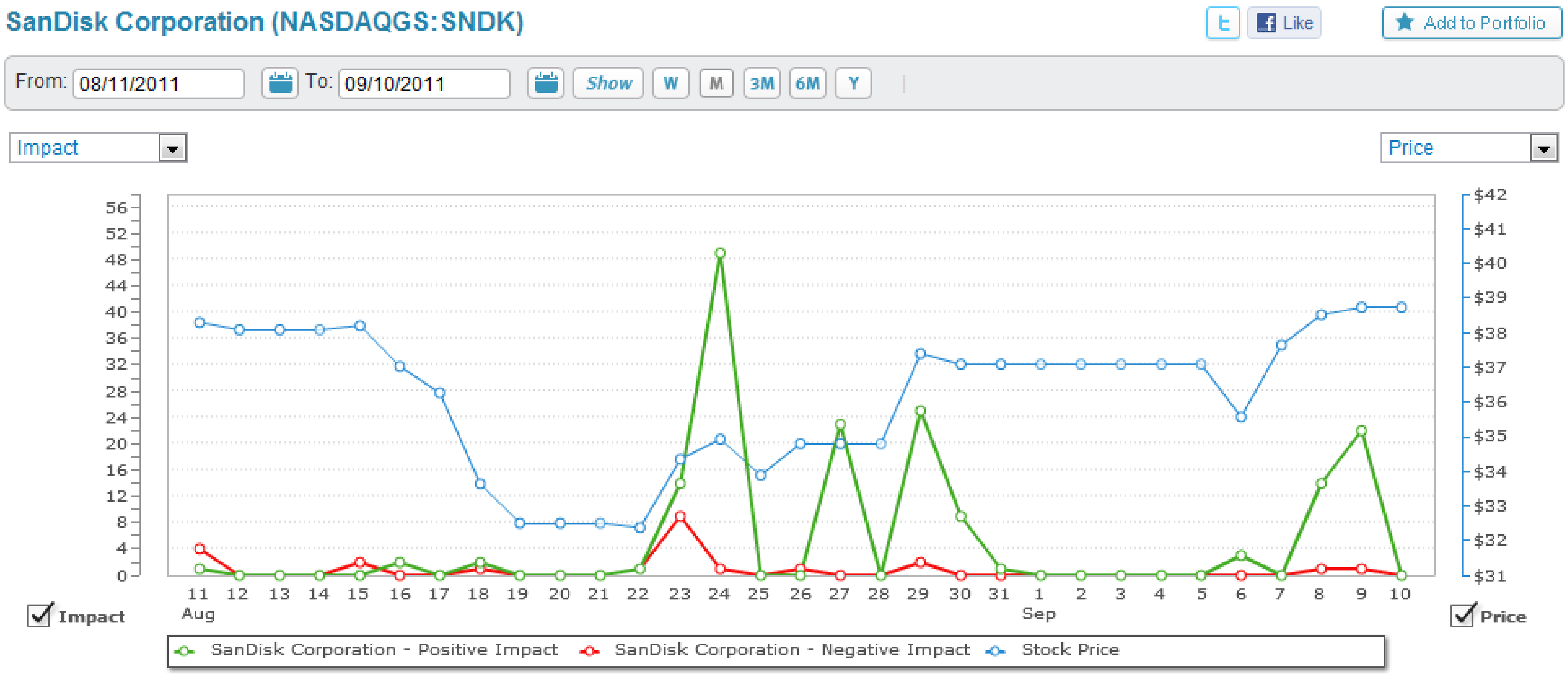
MOS



Mosaic Company The (NYSE:MOS)

From: 05/27/2012 To: 06/26/2012

SNDK



SanDisk Corporation (NASDAQGS:SNDK)

From: 08/11/2011 To: 09/10/2011

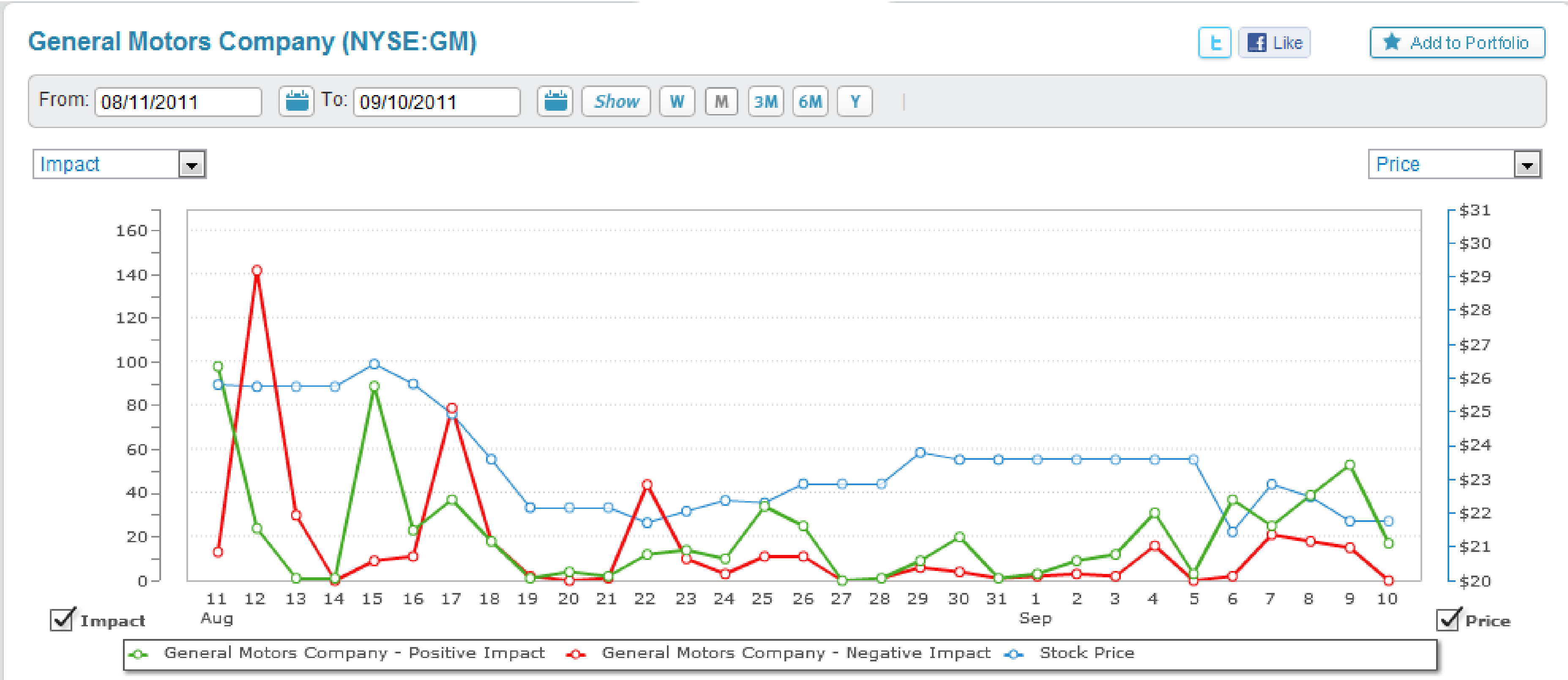
Impact

Price

☑ Impact

☑ Price

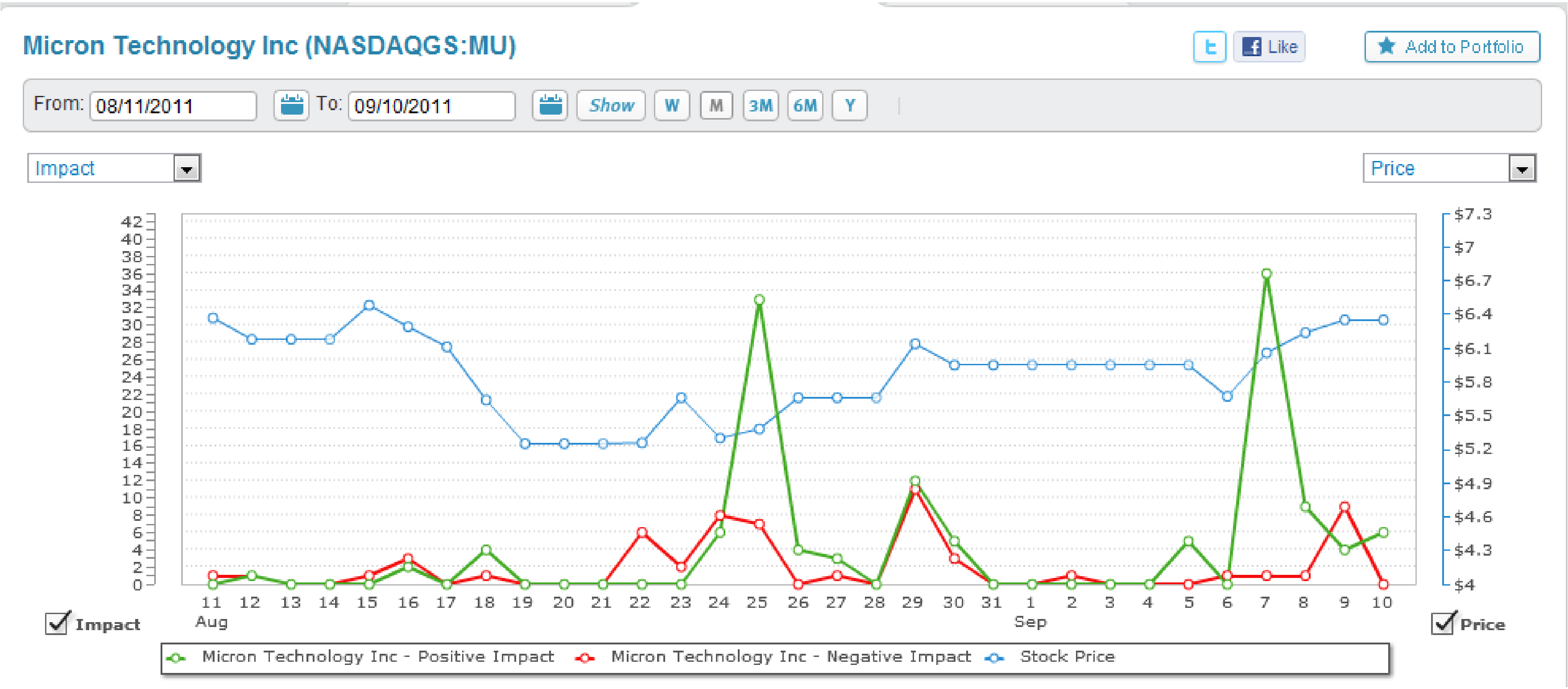
GM



General Motors Company (NYSE:GM)

From: 08/11/2011 To: 09/10/2011

MU



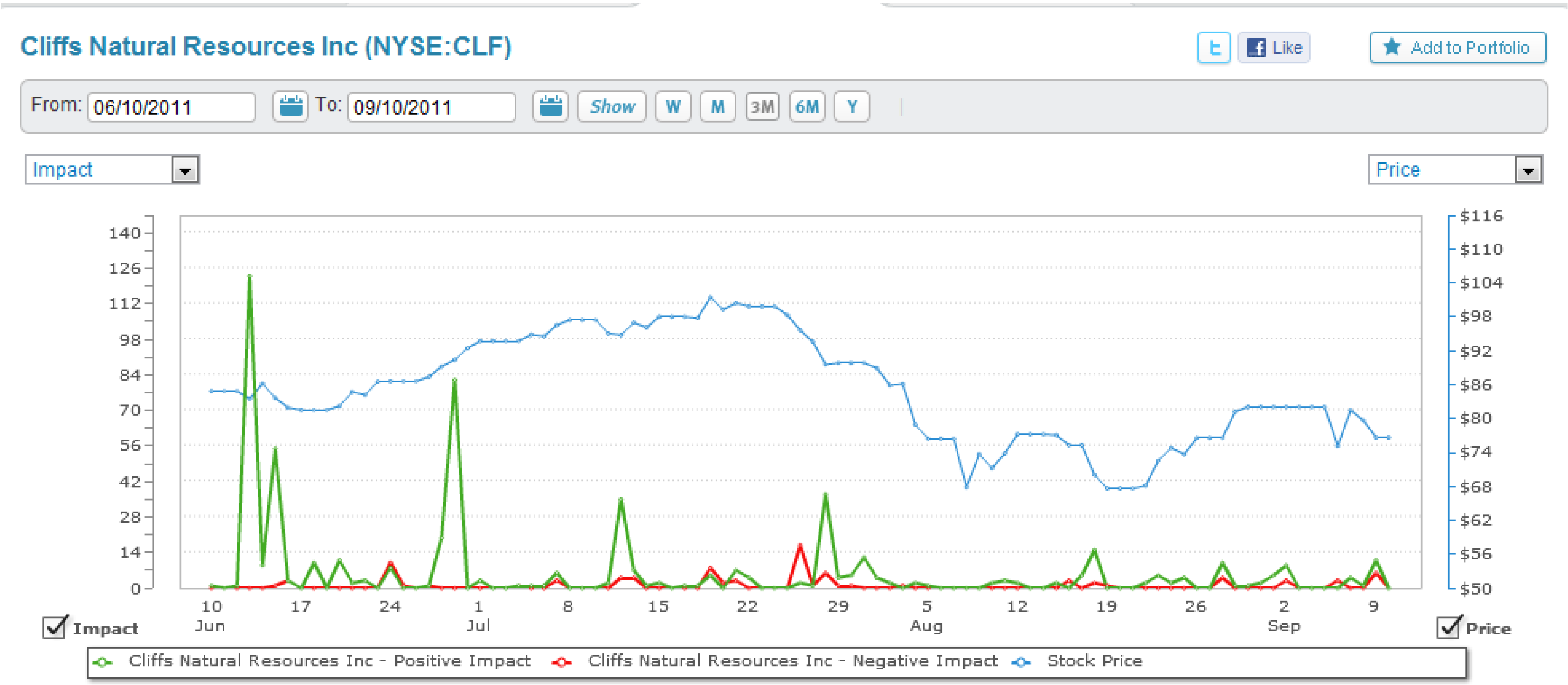
Micron Technology Inc (NASDAQGS:MU)

From: 08/11/2011 To: 09/10/2011

Impact

Price

CLF

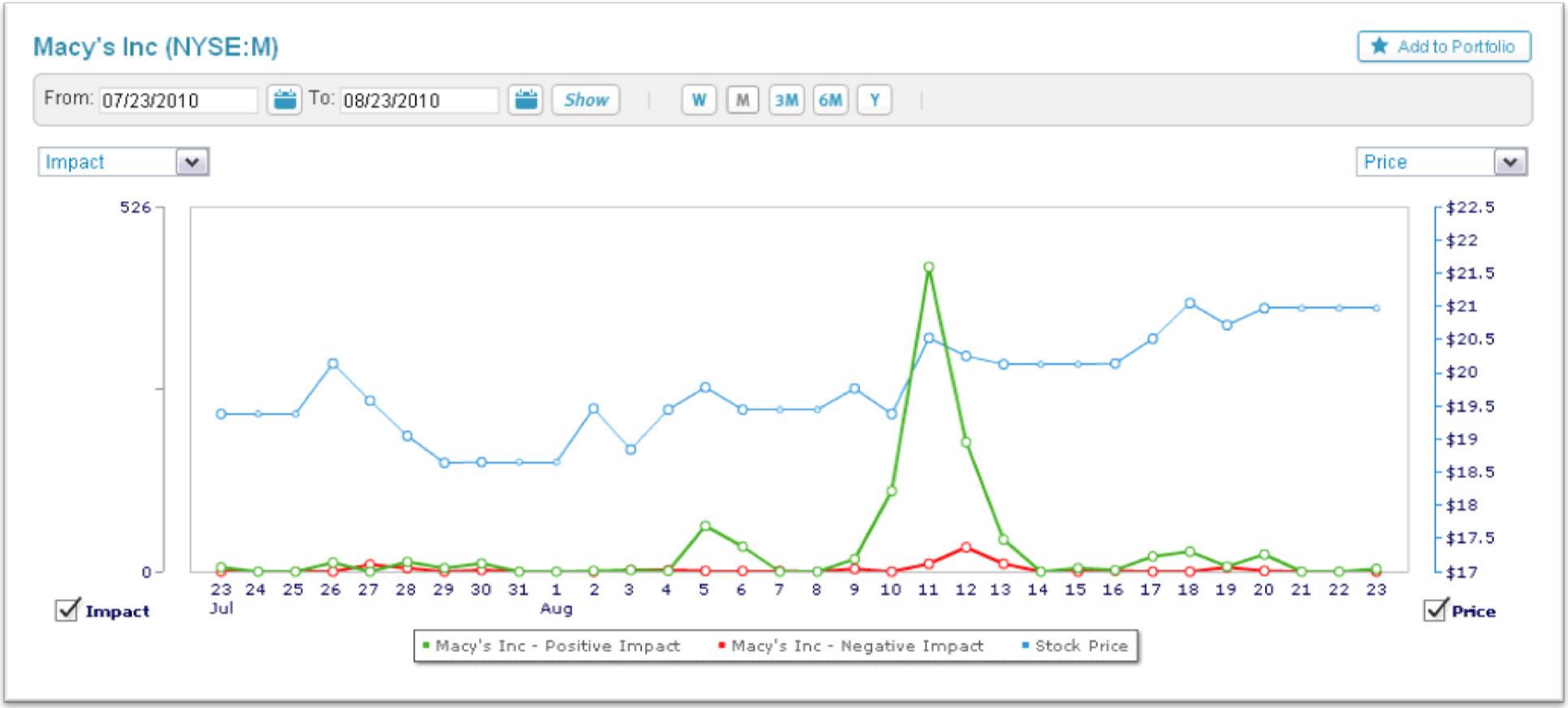


Cliffs Natural Resources Inc (NYSE:CLF)

From: 06/10/2011 To: 09/10/2011

The Stock Sonar
See through the current

Macy's

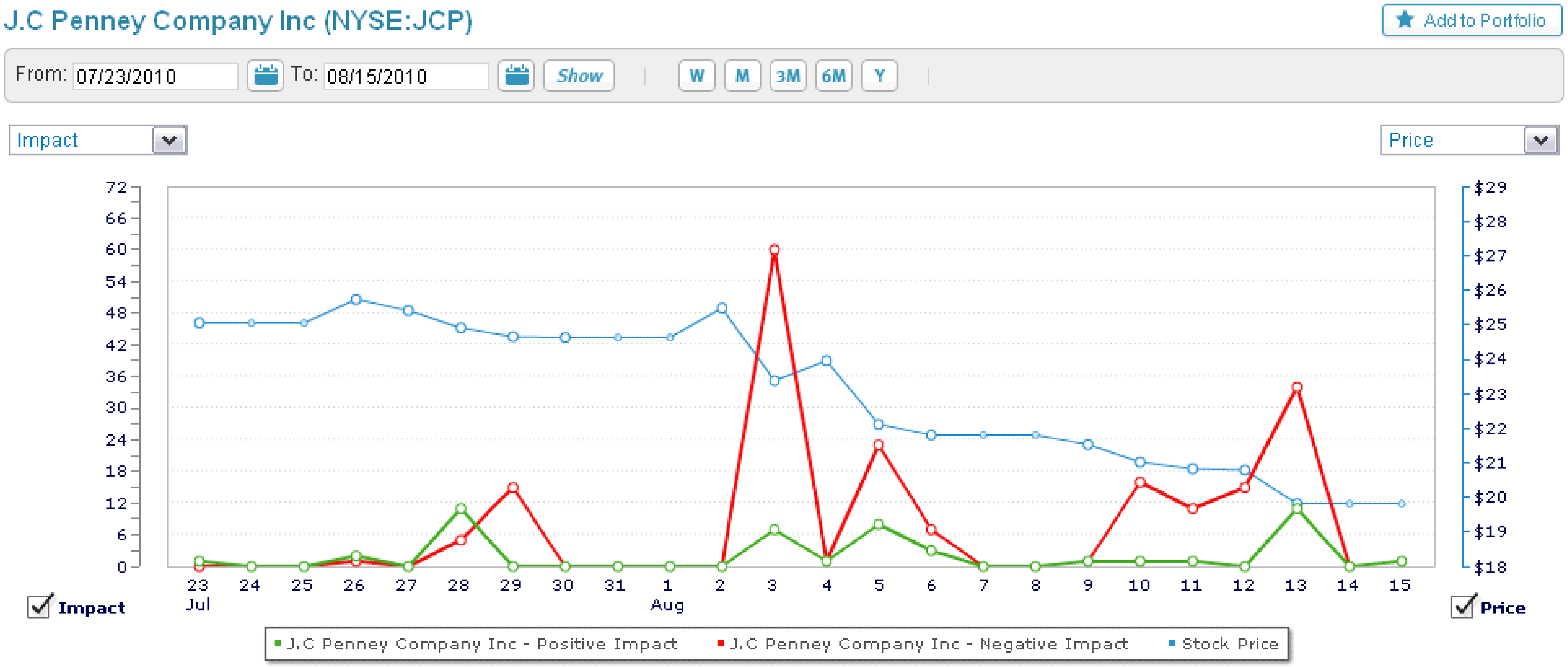


Macy's Inc (NYSE:M)

From: 07/23/2010 To: 08/23/2010

🟩 Macy's Inc - Positive Impact 🟥 Macy's Inc - Negative Impact 🟦 Stock Price

JC Penny



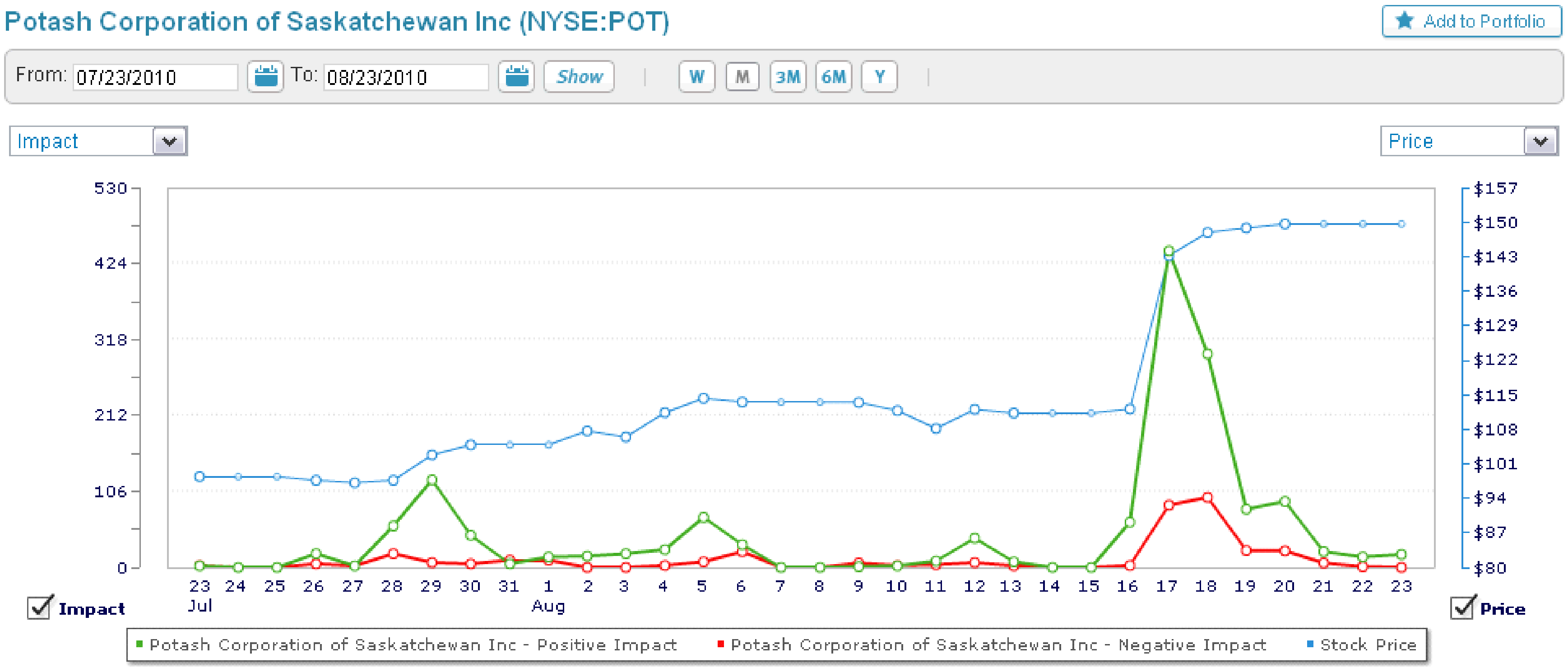
J.C Penney Company Inc (NYSE:JCP)

From: 07/23/2010 To: 08/15/2010

Impact

Price

POT



Potash Corporation of Saskatchewan Inc (NYSE:POT)

From: 07/23/2010 To: 08/23/2010

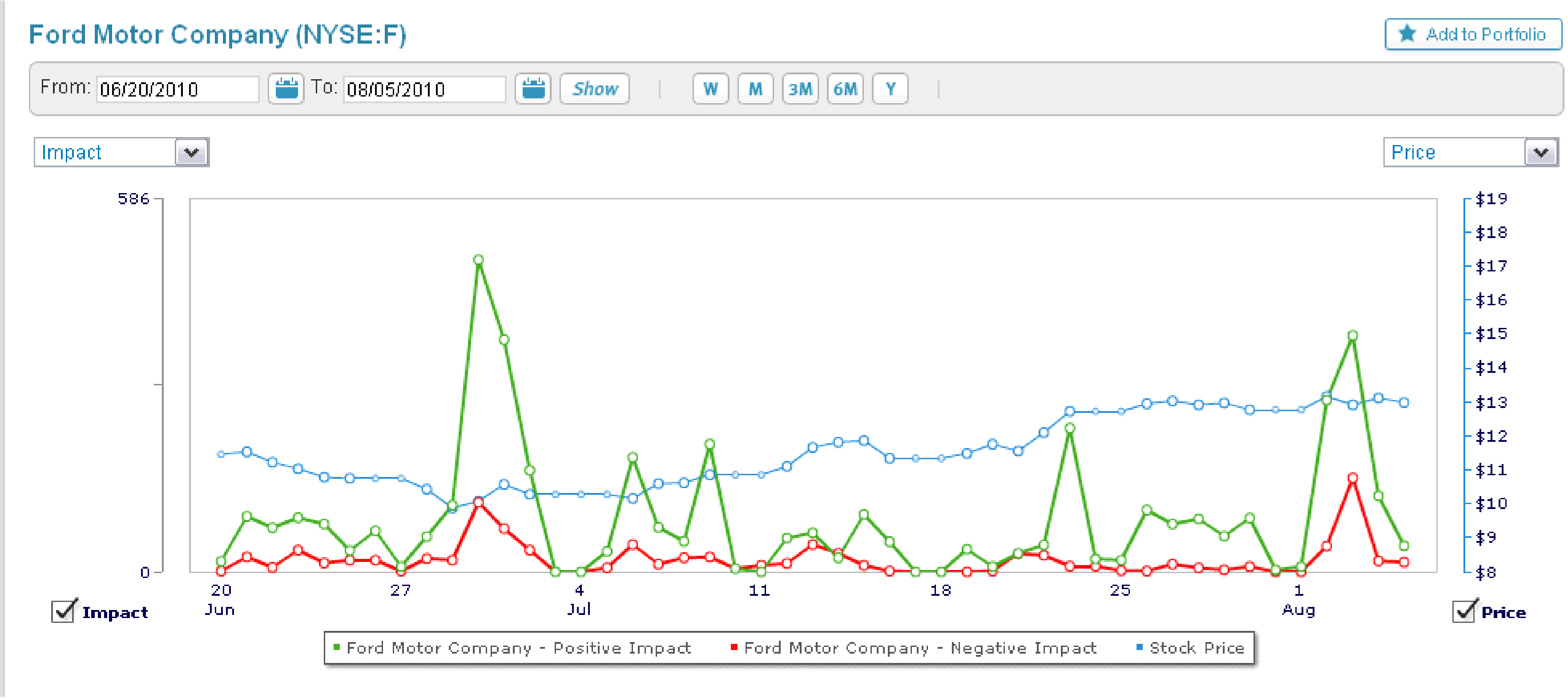
Impact

Price

☑ Impact

☑ Price

Ford



Ford Motor Company (NYSE:F)

From: 06/20/2010 To: 08/05/2010

Impact

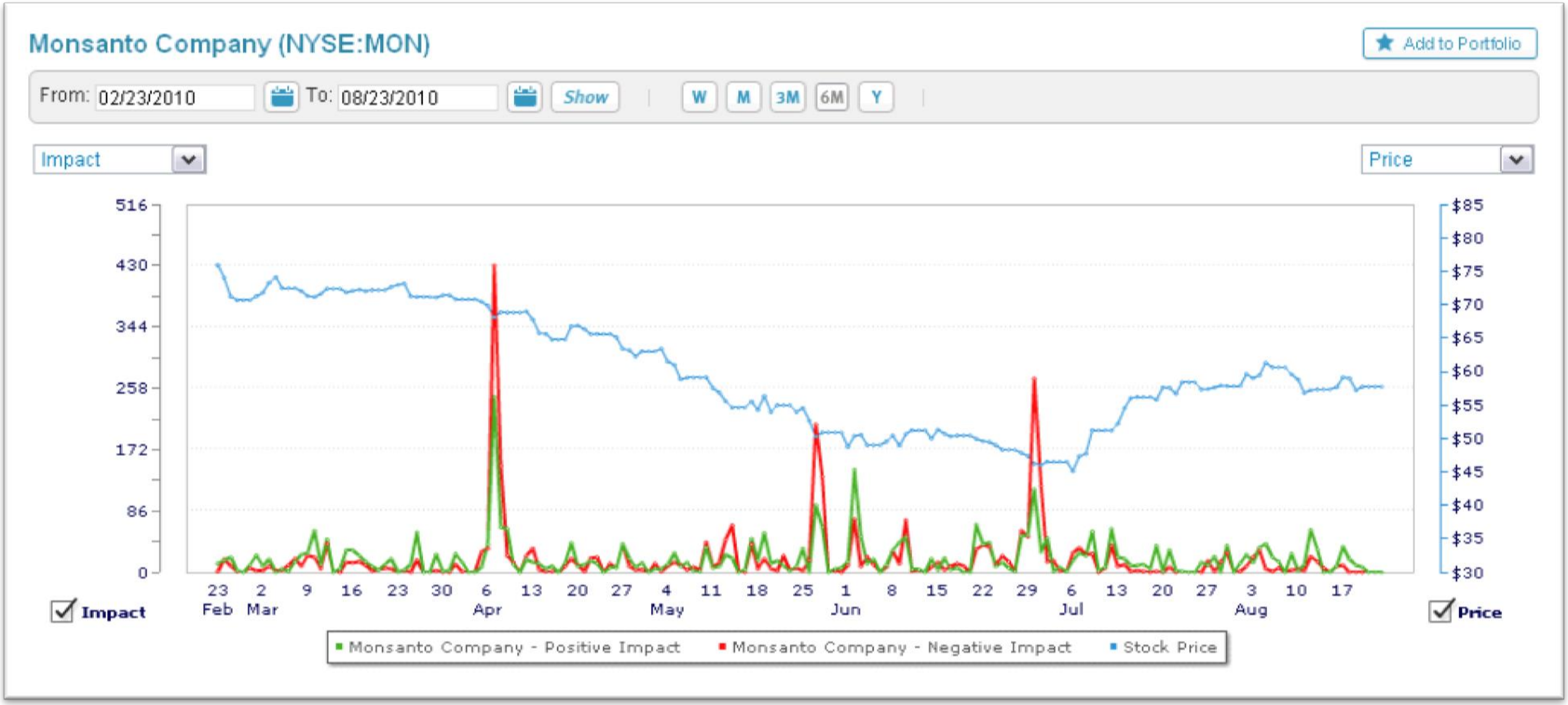
Price

☑ Impact

☑ Price

The Stock Sonar
See through the current

Monsanto



Monsanto Company (NYSE:MON)

From: 02/23/2010 To: 08/23/2010

Impact

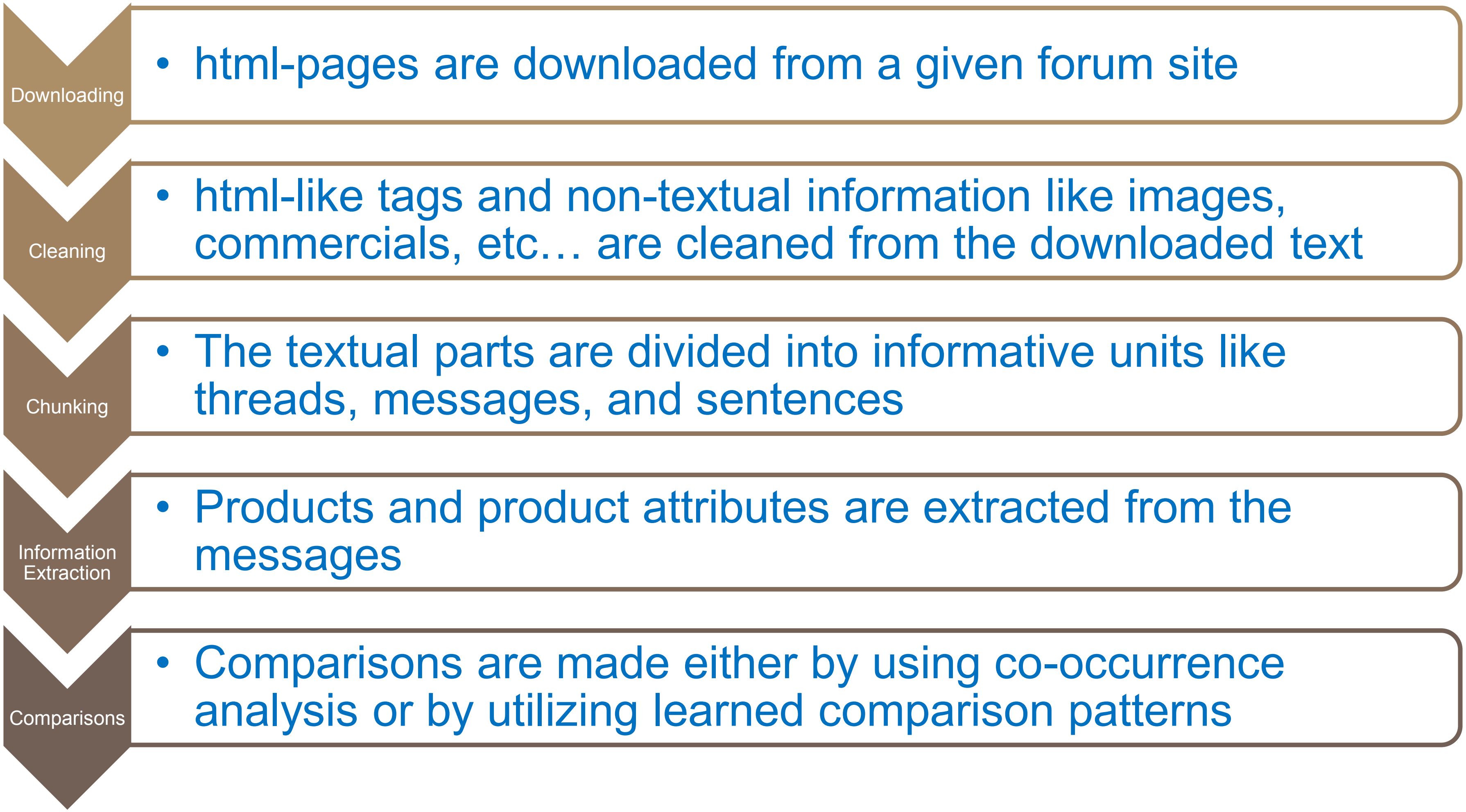
Price

☑ Impact

☑ Price

Mining Medical User Forums

The Text Mining Process



Downloading

- html-pages are downloaded from a given forum site

Cleaning

- html-like tags and non-textual information like images, commercials, etc… are cleaned from the downloaded text

Chunking

- The textual parts are divided into informative units like threads, messages, and sentences

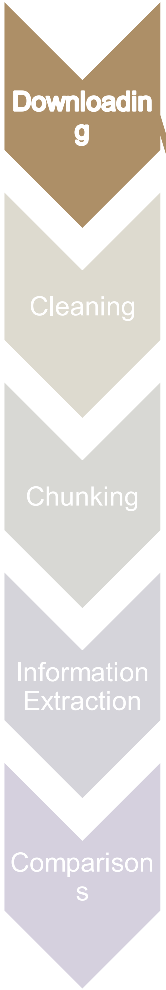
Information Extraction

- Products and product attributes are extracted from the messages

Comparisons

- Comparisons are made either by using co-occurrence analysis or by utilizing learned comparison patterns

The Text Mining Process



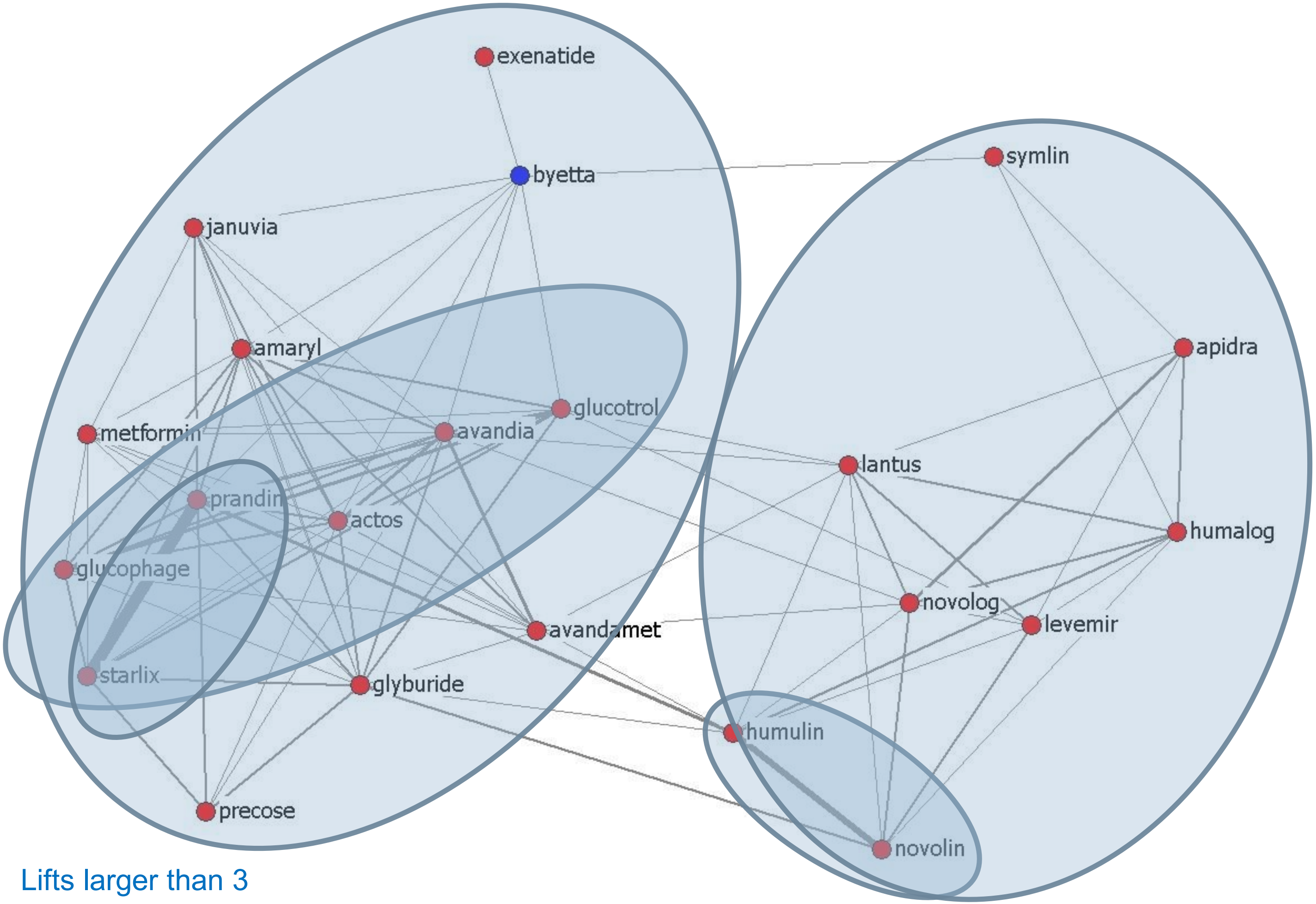
We downloaded messages from 5 different consumer forums

- diabetesforums.com
- healthboards.com
- forum.lowcarber.org
- diabetes.blog.com**
- diabetesdaily.com

** Messages in Diabets.blog.com were focused mainly on Byetta

Drug Analysis

Drug Co-Occurrence - Spring Graph – Perceptual Map



Several Pockets of drugs that were mentioned frequently together in a message were identified

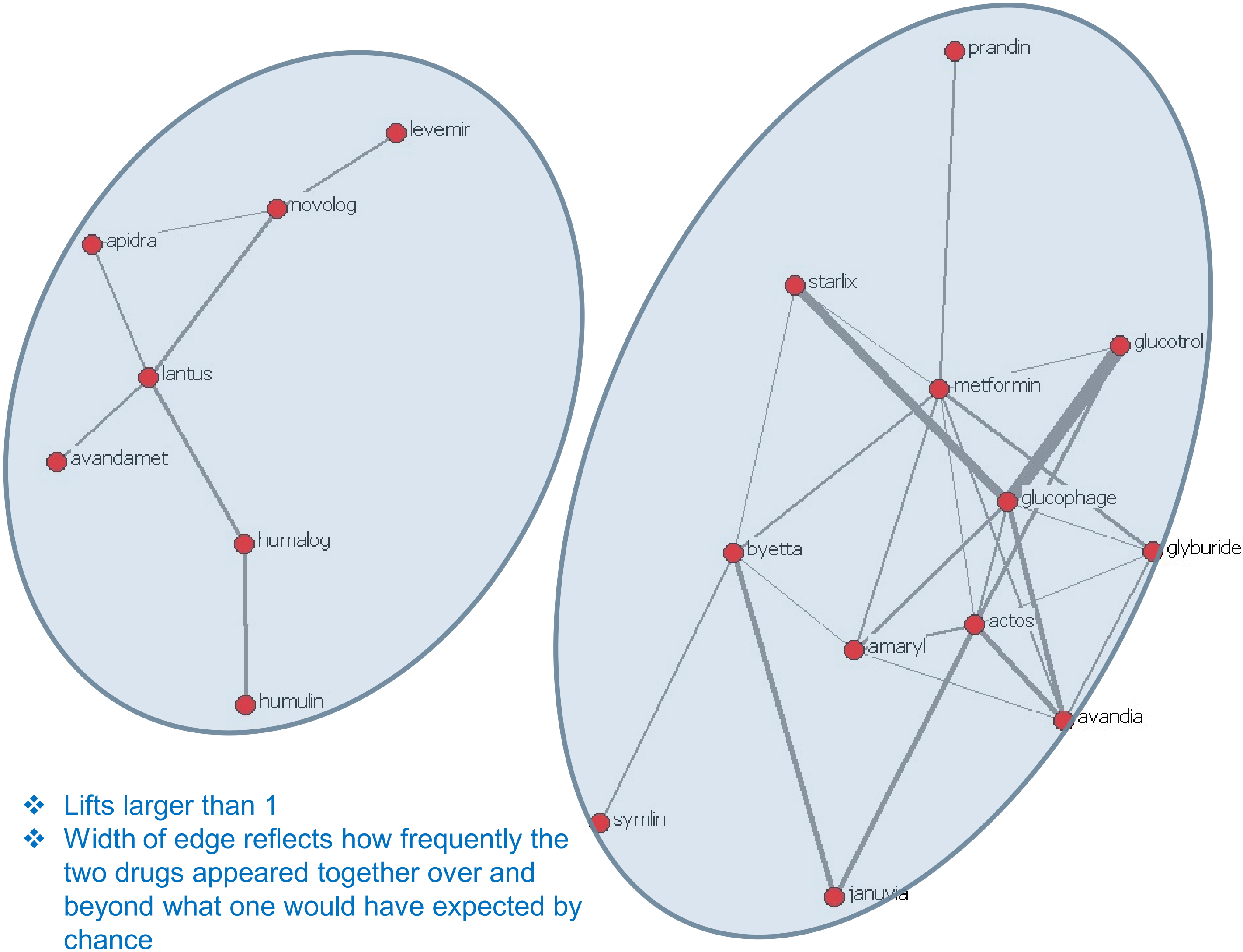
Byetta was mentioned frequently with:

- Glucotrol
- Januvia
- Amaryl
- Actos
- Avandia
- Prandin
- Symlin

- Lifts larger than 3
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Drug Usage Analysis

Drug Co-taking – Drugs mentioned as "Taken Together"



There are two main clusters of drugs that are mentioned as "taken together"

Byetta was mentioned as "taken together" with:

- Januvia
- Symlin
- Metformin
- Amaryl
- Starlix

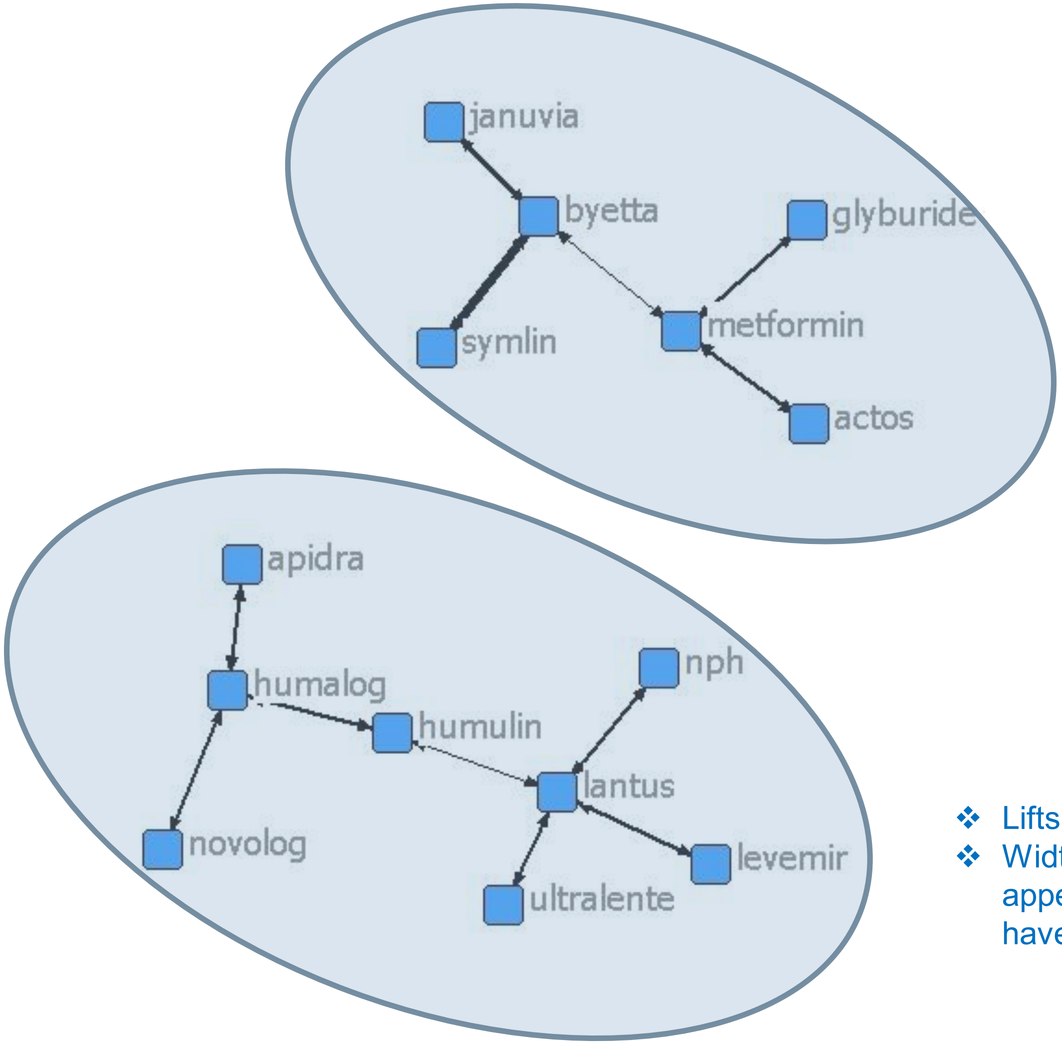
Pairs of drugs that are taken frequently together include:

- Glucotrol--Glucophage
- Glucophage--Stralix
- Byetta--Januvia
- Avandia--Actos
- Glucophage--Avandia

- ❖ Lifts larger than 1
- ❖ Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Drug Usage Analysis

Drug Switching – Drugs mentioned as “Switched” to and from



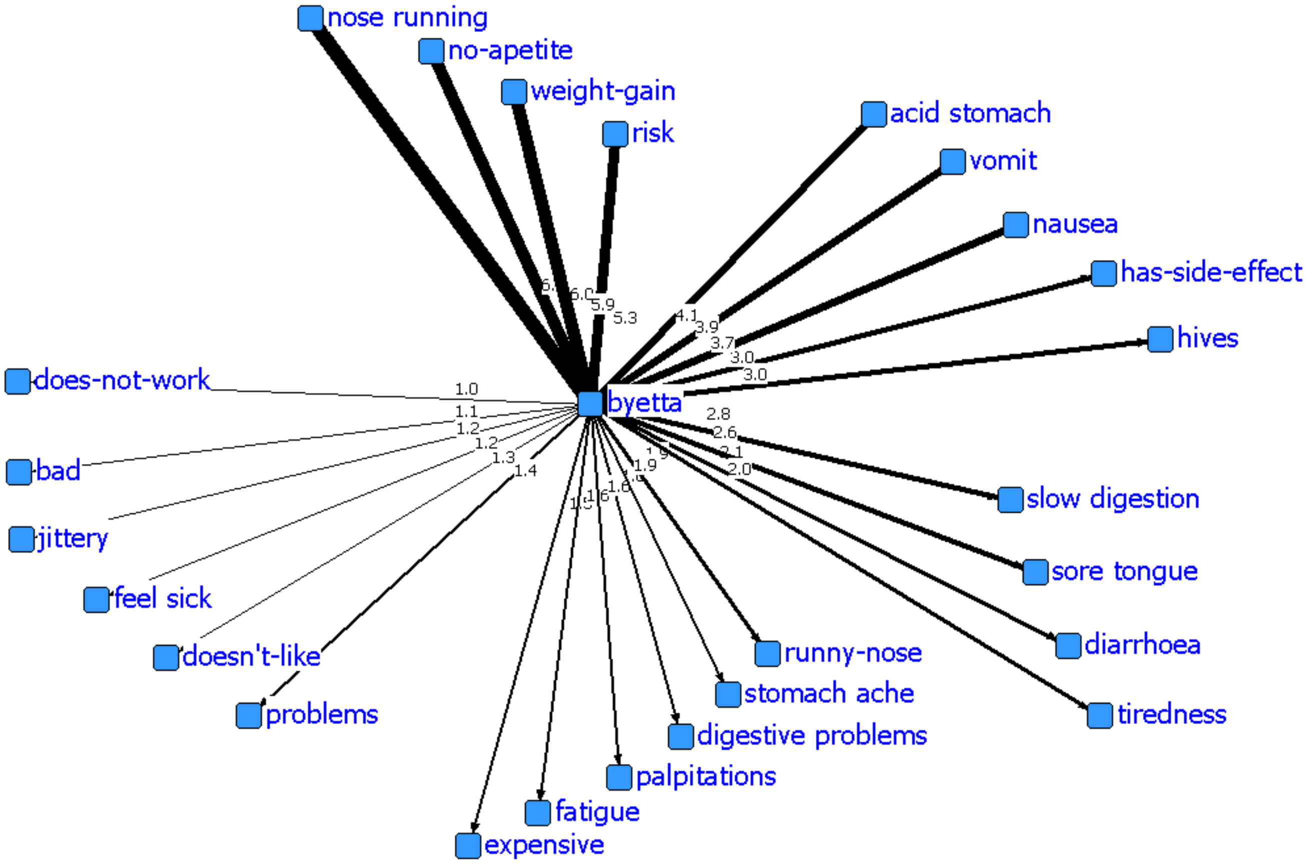
There are two main clusters of diabetes drugs within which consumers mentioned frequently that they “switched” from one drug to another

Byetta was mentioned as “switched” to and from:
- Symlin
- Januvia
- Metformin

- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Drug Terms Analysis

Byetta - Side Effects Analysis

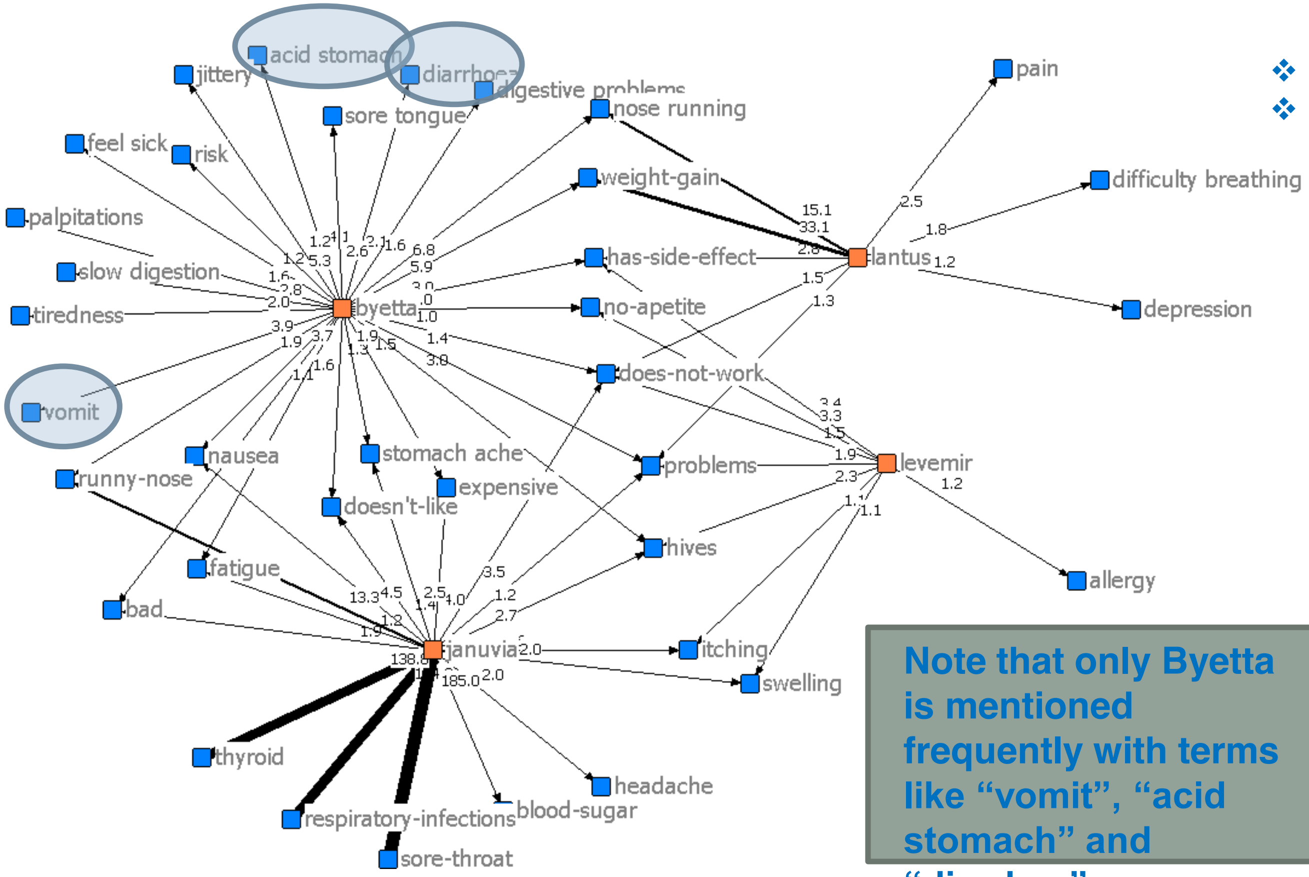


Byetta appeared much more than chance with the following side effects:

- “Nose running” or “runny nose”
- “No appetite”
- “Weight gain”
- “Acid stomach”
- “Vomit”
- “Nausea”
- “Hives”

Drug Terms Analysis

Drug Comparisons on Side Effects



- ❖ Lifts larger than 1
- ❖ Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Note that only Byetta is mentioned frequently with terms like “vomit”, “acid stomach” and “diarrhea”

Byetta shares with Januvia the side effects:
- Runny nose
- Nausea
- Stomach ache
- Hives

Byetta shares with Levemir the side effects:
- No appetite
- Hives

Byetta shares with Lantus the side effects:
- Nose running
- Weight gain

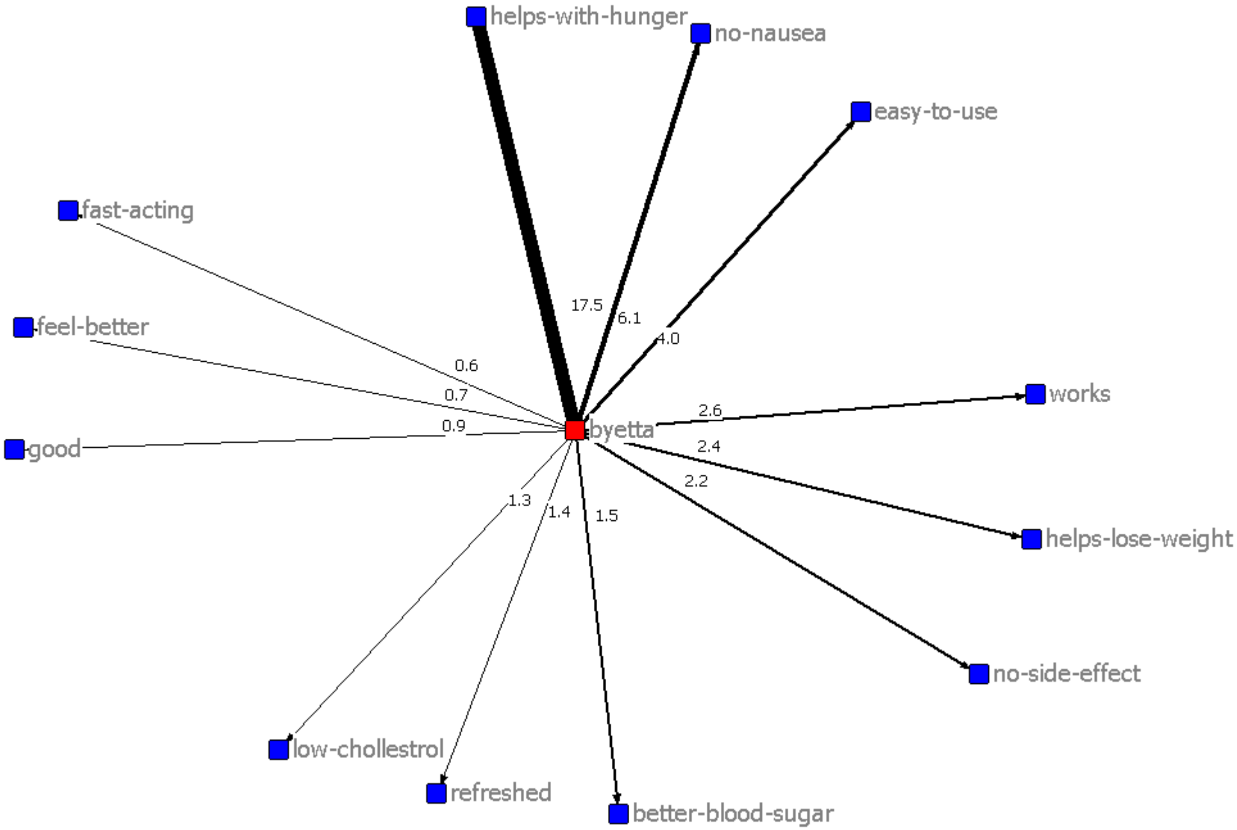
The main side effects discussed with Januvia:
- Thyroid
- Respiratory infections
- Sore throat

The main side effects discussed with Levemir :
- No appetite
- Hives

The main side effects discussed with Lantus:
- Weight gain
- Nose running
- Pain

Drug Terms Analysis

Byetta – Positive Sentiments

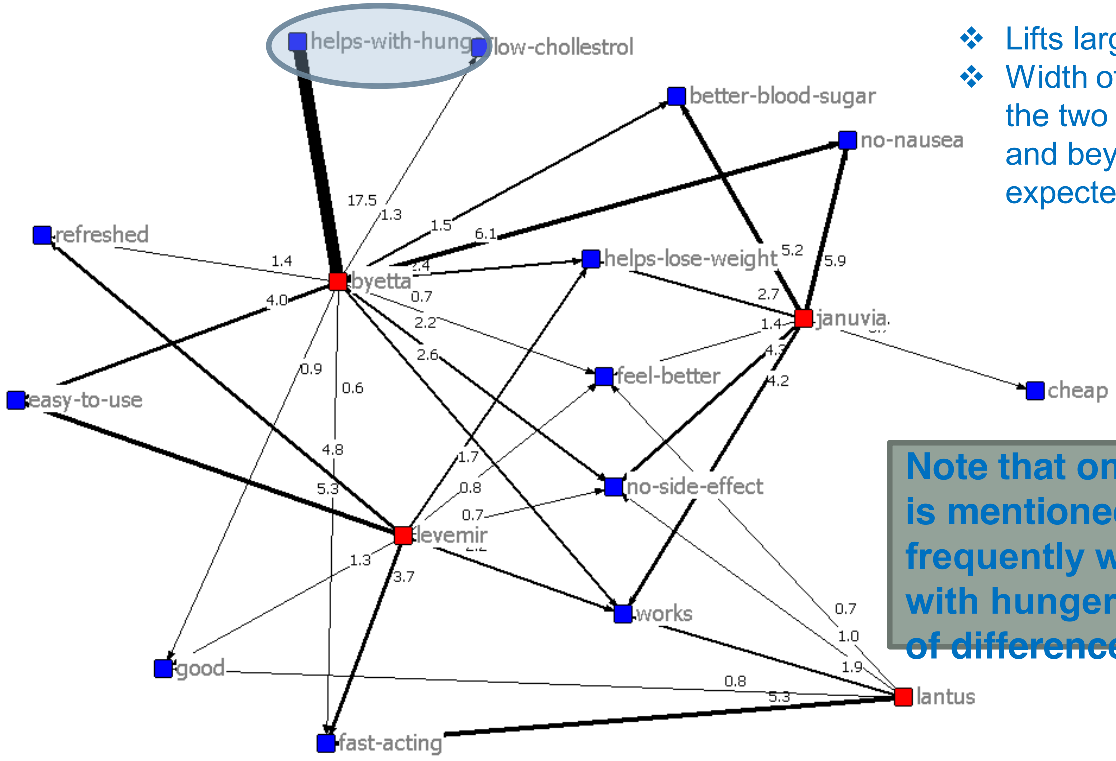


Byetta appeared much more than chance (lift>2) with the following positive sentiments:

- "Helps with hunger"
- "No nausea"
- "Easy to use"
- "Works"
- "Helps losing weight"
- "No side effects"

Drug Terms Analysis

Drug Comparisons on Positive Sentiments



- ❖ Lifts larger than 0.5
- ❖ Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Note that only Byetta is mentioned frequently with "helps with hunger" (point of difference)

The main positive sentiments discussed with <u>Januvia:</u>

- "No nausea"
- "Better blood sugar"
- "Works"
- "No side effects"

The main positive sentiments discussed with <u>Levemir</u> :

- "Easy to use"
- "Fast acting"

The main positive sentiments discussed with <u>Lantus:</u>

- "Fast acting"
- "Works"

Byetta shares with <u>Januvia</u>:

- "Better blood sugar"
- "No nausea"
- "Helps lose weight"
- "No side effects"
- "Works"

Byetta shares with <u>Levemir</u>:

- "Easy to use"
- "Helps lose weight"
- "No side effects"
- "Works"

Byetta shares with <u>Lantus</u>:

- "Easy to use"
- "No side effects"
- "Works"

Example of Unsupervised IE Results

Actos; weight gain (40 (P: 38, N: 2))

Rel_take_DRU_has_SYM(DRUG, SYMPTOM)

Negative (1)

I've been taking 15 mg of **Actos** for just over a year now and so far (knock on wood) I haven't had the weight gain that some others have reported as a side effect.

Positive (8)

I also have read here about some of you who have been on the **Actos** and the weight gain you had experienced.

We saw an endo because of all of the weight gain and side effects from taking **actos**.

He was on **Actos** but went off of it because of weight gain and stomach bloating.

I really don't want to go back on **Actos** because of weight gain/fluid retention.

My doctor wanted me to start **Actos** for awhile, until the Byetta kicks in, but I stopped **Actos** in the first place because of weight gain and I said no to restarting that.

I started taking **Actos** first on May 2, 2007 and I started Metformin 3 weeks later I can not take the Spironolactone till Aug but I have noticed that I have gained weight with these 2 drugs instead of losing and I got a treadmill and do 30 min every morning when I get up and lately I have been doing 30 min at night too because of the weight gain.

I have experienced weight gain as well and i am on **Actos** and insulin and glucophage.

I guess that everything comes with a price, but I'm wondering if most folks who have tried **Actos** have experienced weight gain and the other side effects (edema, headaches, nausea, fatigue, etc.).

Rel_SYMofDRU(SYMPTOM, DRUG)

Positive (5)

I do notice that it increases my hunger, so it is possible that **Actos** weight gain issues may be from hunger being stimulated.

I don't think that a lot of us had made the **Actos** induced weight gain connection.

One reported side effect of **Actos** is weight gain.

Example of Unsupervised IE Results – cont'

Rel_cause_DRUvSYM(DRUG, SYMPTOM)

Negative (1)

Actos hasn't caused any weight gain, I am still losing some.

Positive (25)

I also am on Synthroid, Atenolol, Diovan, Lotrel, Lexapro, Vitorin and Prilosec OTC. I didn't realize that **Actos** can cause a weight gain

as I had never read it as a side effect; however, after reading all of the

comments on this site, I now know why my weight has increased over the past few months since taking on it.

I don't take any oral meds, but from what I have read here, **Actos** causes weight gain because of water retention.

why does the endo think you're a type 1? oral meds are usually given only to type 2's,as type 2's have insulin resistance. oral meds

treat the insulin resistance. type 1's require insulin.....i take actoplus met- which is actos and metformin.actos is like avandia and

i've had no heart issues.....tho-avandia and **actos** can cause weight gain....take care,trish

Actos causes edema and weight gain also.

Actos can cause weight gain (so can Avandia, it's cousin)

Now I have started to see a lot of reports of **Actos** causing weight gain, among other things.

for the record, **Actos** can, and does, cause weight gain/water retention.

I'm on both - what did you hate about Metformin? (**Actos** causes weight gain, metformin weight loss)

Also I hear that the **Actos** causes weight gain, so now I am afraid the new pill will cause me to gain weight.

I'm type 1 so only on insulin, but I have heard that **Actos** can cause weight gain.

Avandia & **Actos**, especially in combination with insulin, causes fluid retention and/or fat weight gain.

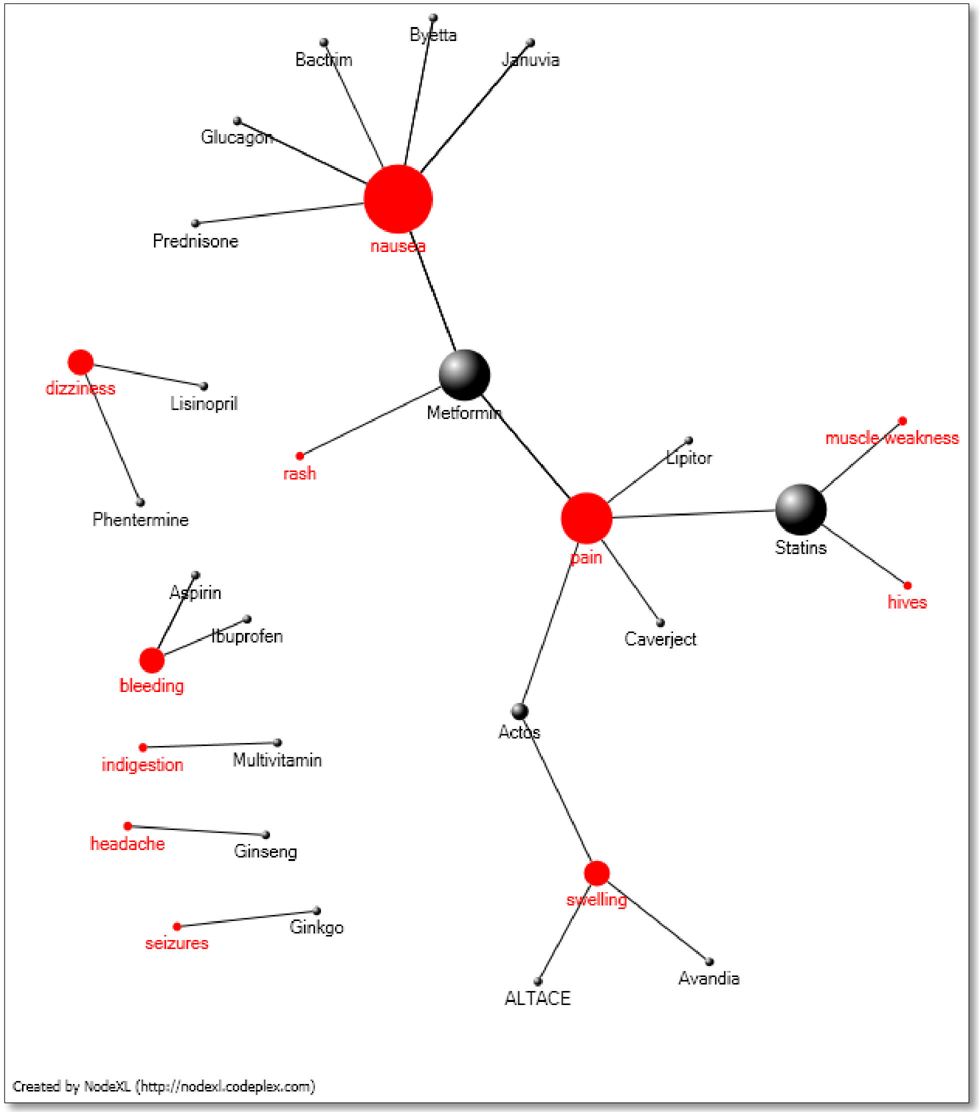
My endocrinologist warned me that **Actos** can cause significant weight gain.

Actos caused weight gain and fluid retention in my chest.

Metformin causes weight loss, Avandia and **Actos** causes the birth of new fat cells and weight gain.

……

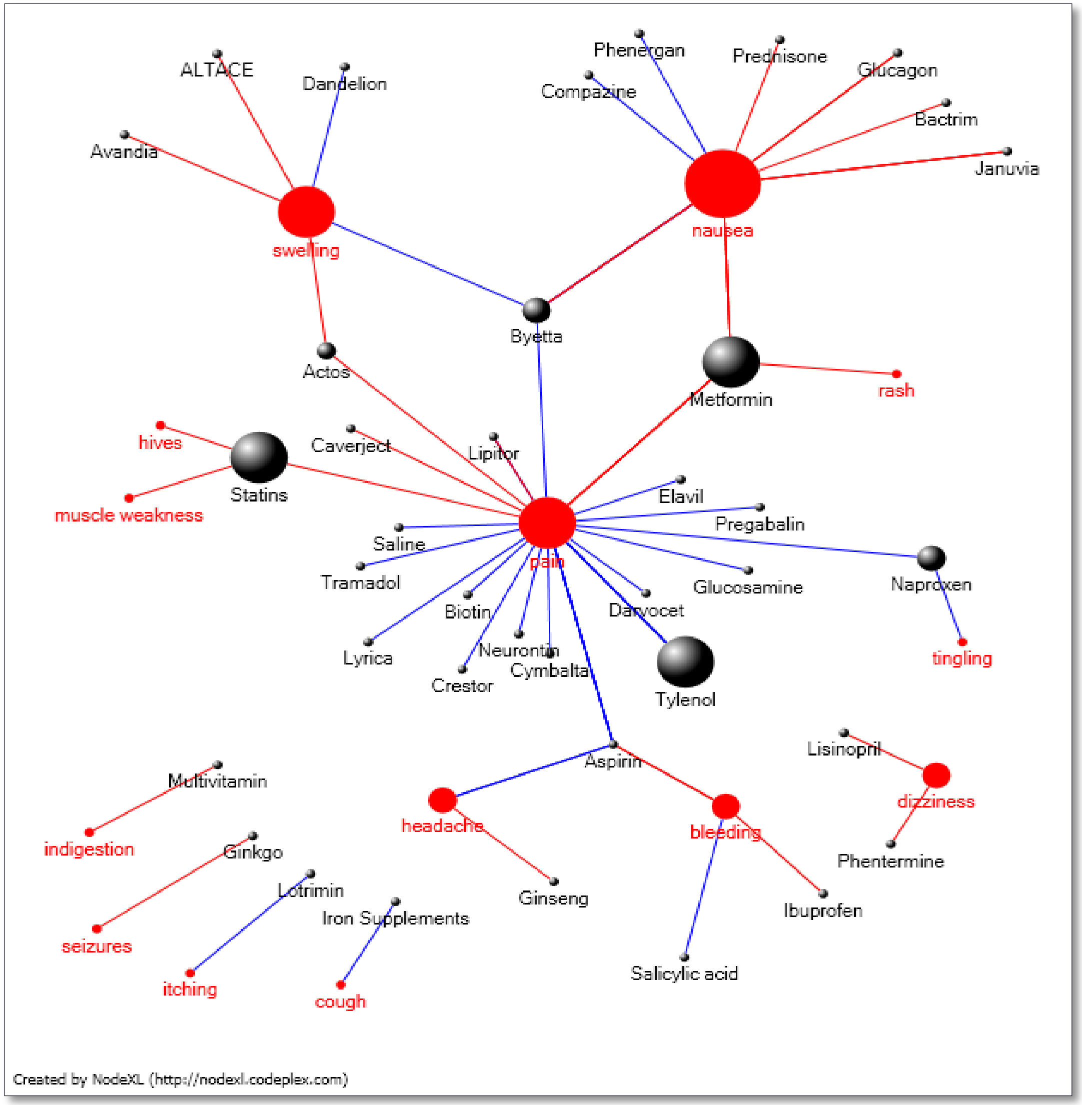
Side Effects



Side Effects and Remedies

Red lines – side effects/symptoms

Blue lines - Remedies



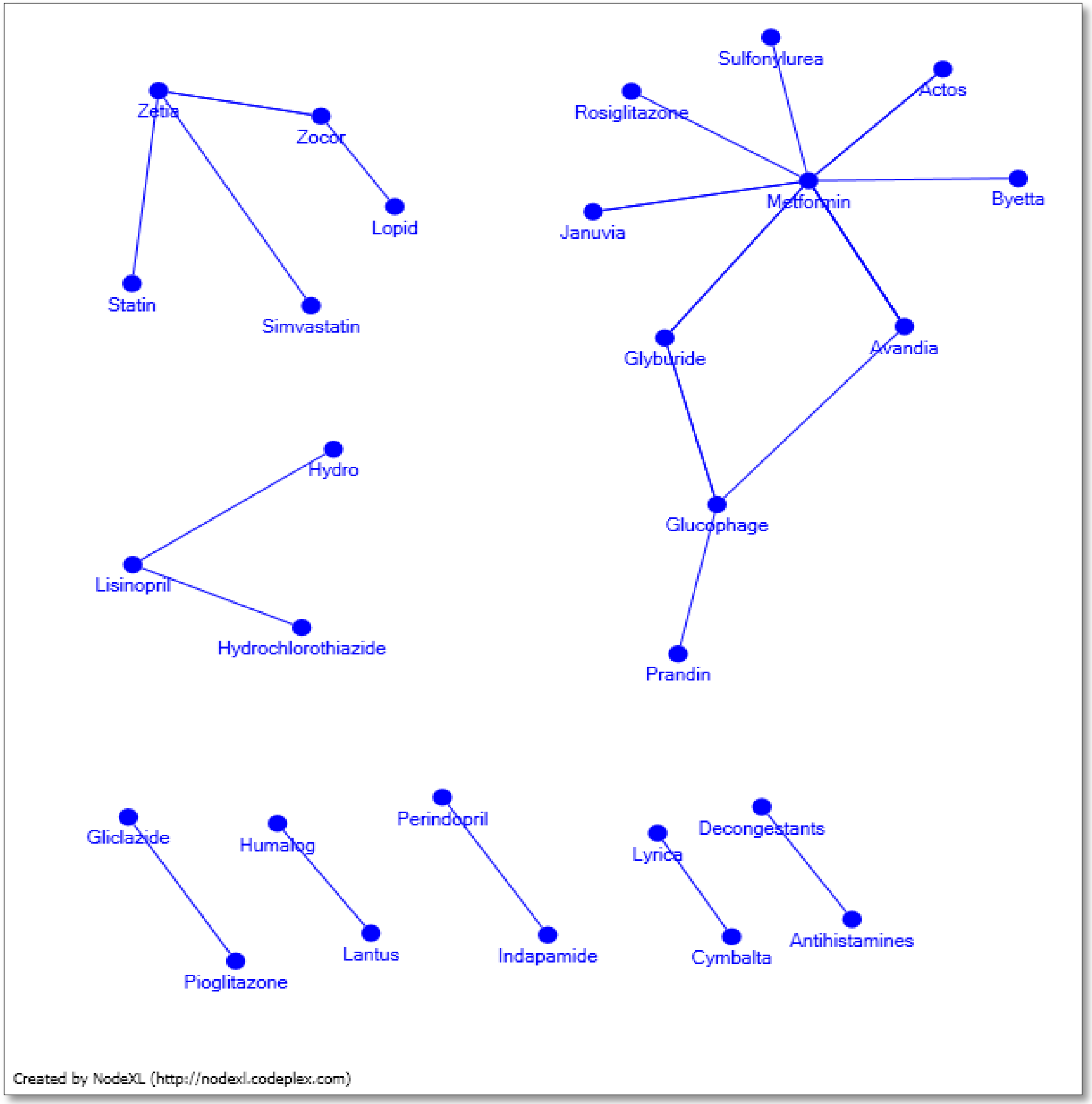
Created by NodeXL (http://nodexl.codeplex.com)

See what causes symptoms and what relieves them

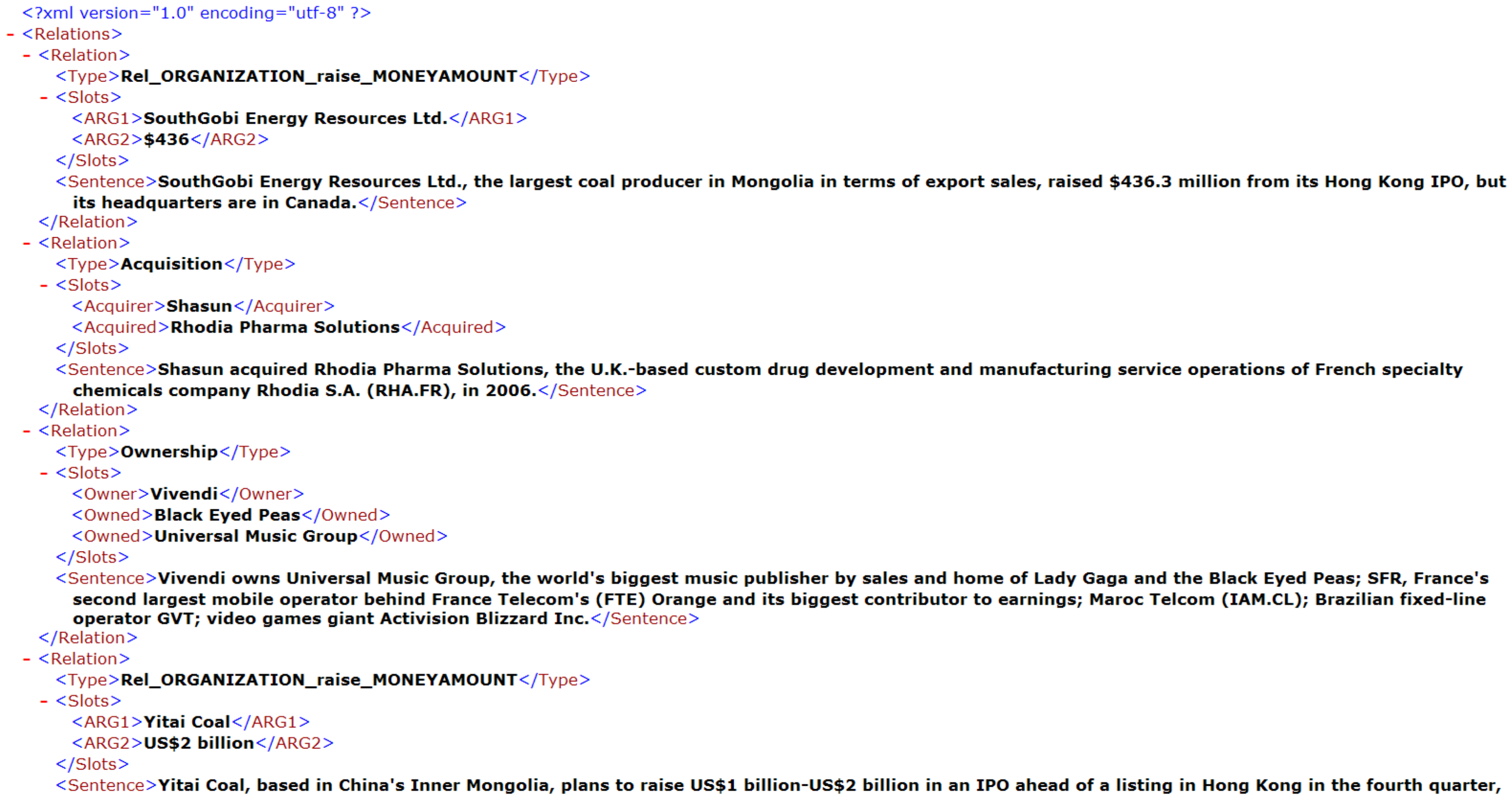
See what positive and negative effects a drug has

See which symptoms are most complained about

Drugs Taken in Combination



XML output of VC



```
  <?xml version="1.0" encoding="utf-8" ?>
- <Relations>
  - <Relation>
      <Type>Rel_ORGANIZATION_raise_MONEYAMOUNT</Type>
    - <Slots>
        <ARG1>SouthGobi Energy Resources Ltd.</ARG1>
        <ARG2>$436</ARG2>
      </Slots>
      <Sentence>SouthGobi Energy Resources Ltd., the largest coal producer in Mongolia in terms of export sales, raised $436.3 million from its Hong Kong IPO, but
        its headquarters are in Canada.</Sentence>
    </Relation>
  - <Relation>
      <Type>Acquisition</Type>
    - <Slots>
        <Acquirer>Shasun</Acquirer>
        <Acquired>Rhodia Pharma Solutions</Acquired>
      </Slots>
      <Sentence>Shasun acquired Rhodia Pharma Solutions, the U.K.-based custom drug development and manufacturing service operations of French specialty
        chemicals company Rhodia S.A. (RHA.FR), in 2006.</Sentence>
    </Relation>
  - <Relation>
      <Type>Ownership</Type>
    - <Slots>
        <Owner>Vivendi</Owner>
        <Owned>Black Eyed Peas</Owned>
        <Owned>Universal Music Group</Owned>
      </Slots>
      <Sentence>Vivendi owns Universal Music Group, the world's biggest music publisher by sales and home of Lady Gaga and the Black Eyed Peas; SFR, France's
        second largest mobile operator behind France Telecom's (FTE) Orange and its biggest contributor to earnings; Maroc Telcom (IAM.CL); Brazilian fixed-line
        operator GVT; video games giant Activision Blizzard Inc.</Sentence>
    </Relation>
  - <Relation>
      <Type>Rel_ORGANIZATION_raise_MONEYAMOUNT</Type>
    - <Slots>
        <ARG1>Yitai Coal</ARG1>
        <ARG2>US$2 billion</ARG2>
      </Slots>
      <Sentence>Yitai Coal, based in China's Inner Mongolia, plans to raise US$1 billion-US$2 billion in an IPO ahead of a listing in Hong Kong in the fourth quarter,
```

USING MULTI-VIEW LEARNING TO IMPROVE DETECTION OF INVESTOR SENTIMENTS ON TWITTER

Motivation

- Stocks-related messages on social media, Twitter in particular, have several interesting properties with regards to the sentiment analysis task.
- The analysis is particularly challenging:
  - frequent typos
  - bad grammar
  - idiosyncratic expressions specific to the domain and the media.
- Stocks-related messages almost always refer to the state of specific entities – companies and their stocks – at specific times (times of sending). This state is an objective property and even has a measurable numeric characteristic, namely the stock price.

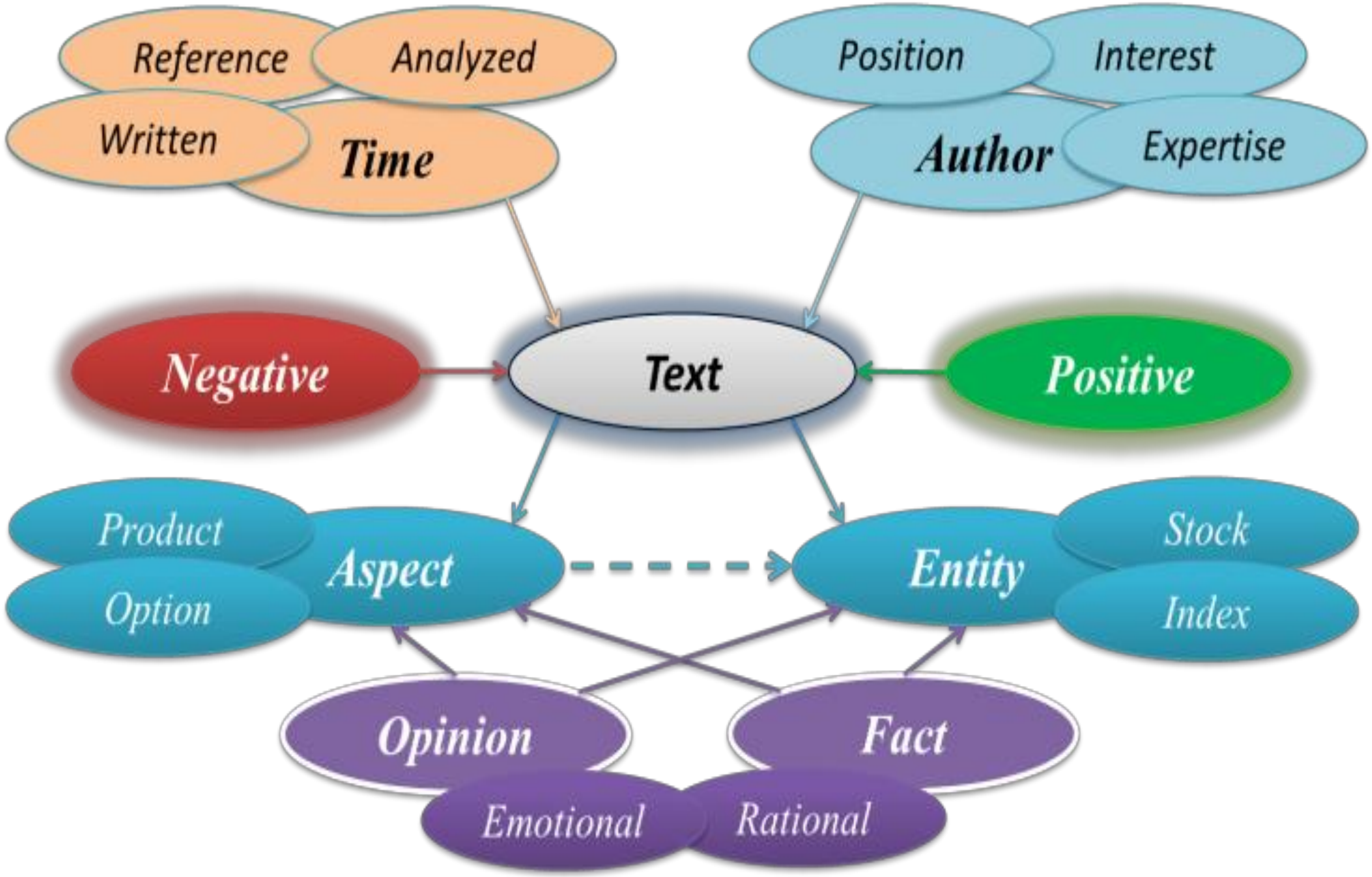
Multi View Learning

- Given a large dataset of twitter messages ("twits"), it is possible to create two separate "views" on the dataset, by analyzing the text of the twits and their external properties separately.
- Using the two views, we can expand the coverage from sentiments detectable using generic sentiment analysis tools to many other twits in the dataset.
- We can use the results to learn new sentiment expressions.

Agenda

1. Complexity and challenges associated with analyzing finance-related message.
2. Related work.
3. Methodological approach for multi-view learning.
4. The SA system.
5. Experimental design and results.
6. Concludes and Future Research.

Complexity and challenges associated with analyzing finance-related message



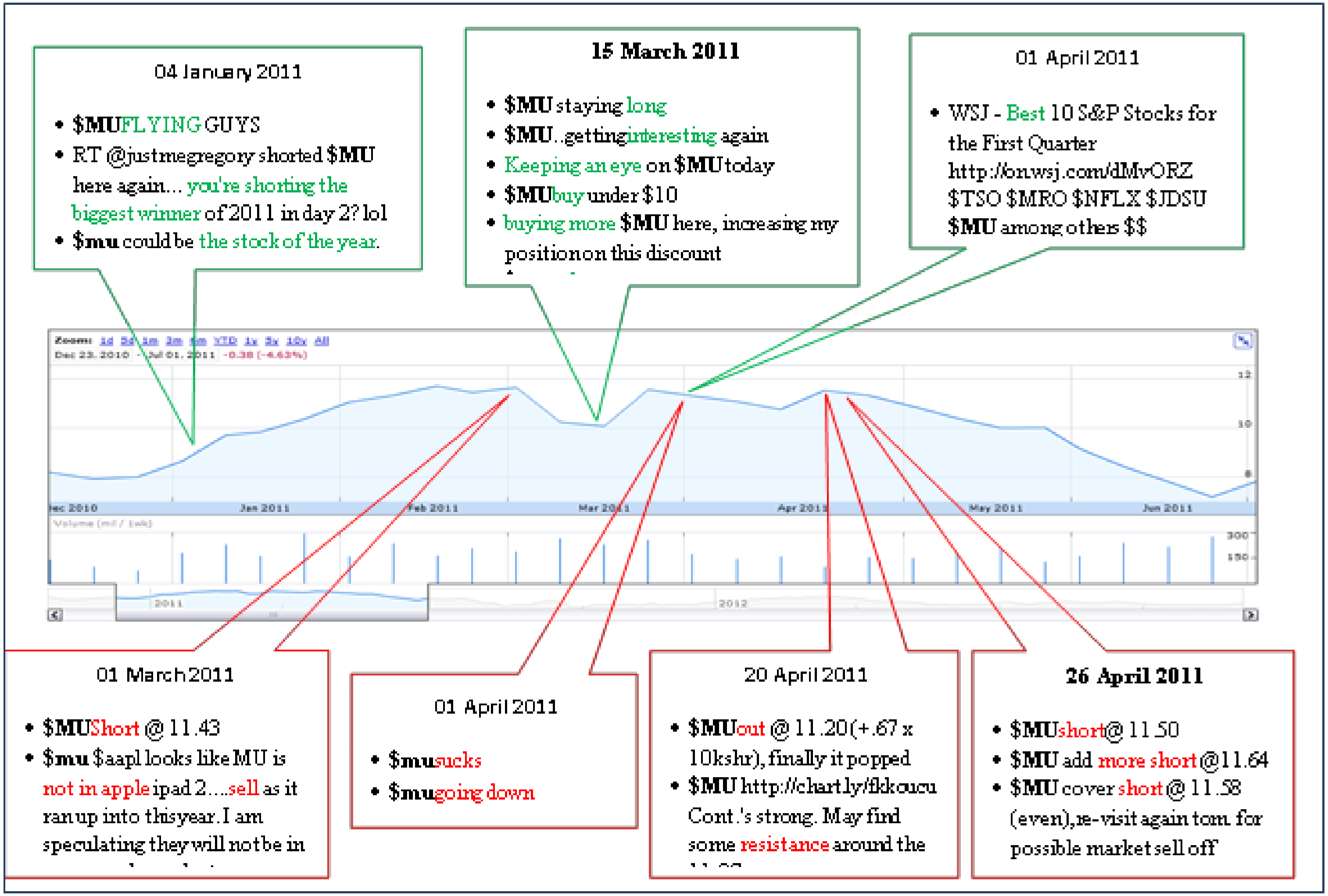
Language Challenges

- Sarcasm - *“It's rumored that Facebook will announce a new kind of free stock dividend tonight, namely CrashVille. $FB $ZNGA #SoCalledEarnings”*
- Price Dependency - *“$SINA watch $57.50 which now is a resistance, $59.21 that will open the door to $62.48 before earning, if not we will break $50”*.
  - If the price at the time of the analysis is 59.21, the sentiment is very positive, as the author believes it can reach 62.48,
  - otherwise, it is negative as the author believes it can drop to $50.

General SA Issues

- Sentiment modifiers (e.g., “highly”)
- Emphasis modifiers (e.g., “mostly”)
- Opposite modifiers(e.g., “far from”)
- Sentiment shifters (“e.g., not”).
- Anaphora resolution is also a challenge when conducting sentiment analysis, although, the short messages on social media often bear little anaphora.

Trading MU (Micron)



04 January 2011

- $MU FLYING GUYS
- RT @justmegregory shorted $MU here again… you're shorting the biggest winner of 2011 in day 2? lol
- $mu could be the stock of the year.

15 March 2011

- $MU staying long
- $MU..getting interesting again
- Keeping an eye on $MU today
- $MU buy under $10
- buying more $MU here, increasing my position on this discount

01 April 2011

- WSJ - Best 10 S&P Stocks for the First Quarter http://on.wsj.com/dMvORZ $TSO $MRO $NFLX $JDSU $MU among others $$

01 March 2011

- $MU Short @ 11.43
- $mu $aapl looks like MU is not in apple ipad 2….sell as it ran up into this year. I am speculating they will not be in

01 April 2011

- $mu sucks
- $mu going down

20 April 2011

- $MU out @ 11.20 (+.67 x 10kshr), finally it popped
- $MU http://chart.ly/fkkoucu Cont.'s strong. May find some resistance around the

26 April 2011

- $MU short @ 11.50
- $MU add more short @ 11.64
- $MU cover short @ 11.58 (even), re-visit again tom. for possible market sell off

Figure 2: Practical illustration

Related work – Text mining and Finance

- **Engelberg (2008)**
    - Found that qualitative earnings information has additional **predictability for asset prices** beyond the predictability in quantitative information.
- **Loughran and McDonald (2011)**
    - Uses textual analysis to examine **the sentiment of corporate 10-K reports**.
- **Feldman at al. (2011)**
    - The Stock Sonar uses state-of-the-art information extraction technologies to analyze **news relating to stocks** and aggregate the sentiments into a score and graphically present the information.
- **Boudoukh et al. (2012)**
    - Found evidence of a **strong relationship between stock price changes and information arrival** by using relevant news articles as a proxy.
    - Information arriving to investors sooner can provide a significant competitive advantage to those investors. Vis-à-vis, it is reasonable to assume that social media response to news is much quicker than professional media.

Related work – Social networks

- **Sprenger & Welpe (2010)**
  - Found that tweet sentiment **associated with abnormal stock returns** and message volume to be a good predictor for next-day trading volume.
- **Bar-haim et al. (2011)**
  - Proposed a general framework for **identifying expert investors**, and used it as a basis for several models that predict stock rise from stock microblogging messages.
- **Bollen, Mao, and Zeng (2011)**
  - Investigated whether measurements of **collective mood** states derived from large-scale Twitter feeds are correlated to the **value of the DJIA over time**.
  - Their results suggested that the accuracy of DJIA predictions can be significantly improved by including specific public mood dimensions.
- **Xu (2012)**
  - Used NLP to get the public sentiment on individual stocks from social media.
  - Found that users' **activity overnight** significantly correlates positively to the stock **trading volume** the next business day.
  - Found **Granger Causality** (in 9 out of 15 stocks studied) between collective sentiments for afterhours and the change of stock price for the next day

Analysis

- One particular point in time that should be noted in the figure is that on April 1st 2011, there were both positive and negative messages about the company.
- However, the positive messages discussed the stock past performance (which was, indeed, positive for the referenced period),
- while the negative messages reflect the investors current and future expectations from the stock (which appears to correlate with the actual stock price's trends in the presented example).

Basic Learning Model

- Assume first, that we have a large corpus $T = \{t_1, t_2, \ldots\}$ of text messages. Each message t has a true polarity $Pol(t) \in POLS=\{POS, NEG, NEUTRAL\}$, which can always be identified (by people) by reading and understanding the text.
- We make further assumption that the polarity of a message is determined by occurrence of various *sentiment expressions* within the message's text.
- The nature of the expressions is not essential at this point. They can be individual words, multi-word sequences, sequences with gaps, syntactic patterns, etc. The important point is that given a message t_i, it must be easy to list all expressions occurring within it. For the purposes of sentiment expression learning, we represent message texts as bags-of-expressions: $t_i = \{w_{i1}, w_{i2}, \ldots\}$.

Observations

- If the true polarity *Pol*(*t*) were known for each message $t \in T$, there would be a natural way to search for sentimental expressions by simply counting their occurrences. However, in a large unlabeled corpus, the true polarities of messages are unknown.
- On the other hand, the polarity of a message is causally-independently influenced by external (relative to the message's text) factors:
  - if a given stock is doing good or bad on a given day, then the polarity of the messages about that stock would tend to be correspondingly positive or negative;
  - messages from the same author about the same stock would tend to have the same polarities; etc.
- Thus, we have two parallel views on a set of twits, which allows multi-view learning.

Parallel Views – Textual View

- Given a large corpus T we first process it with some text-based SA (sentiment analysis) system, which performs as a function $SA : T \rightarrow POLS$, producing a classification $T = T_{SA\text{-}POS} \cup T_{SA\text{-}NEG} \cup T_{SA\text{-}NEUTRAL}$.
- The SA system is assumed to have a relatively high precision for polarized twits, but insufficient recall.
- $T_{SA\text{-}POS}$ and $T_{SA\text{-}NEG}$ generally contain mostly positive and negative messages, respectively
- $T_{SA\text{-}NEUTRAL}$ cannot be assumed to contain only neutral twits. It is also much bigger than the two polarized sets.

Parallel Views : The Second View

- We process the corpus *T* with a feature extractor, which generates a real-valued high-dimensional vector for each message, using any possible property of it that is conditionally independent from its text content given its polarity.
- Using this representation, we train a binary SVM classifier with $T_{SA\text{-}POS}$ and $T_{SA\text{-}NEG}$ as the training data. This classifier then produces a score *f*(*t*) for each message *t* in $T_{SA\text{-}NEUTRAL}$.
- The significant properties of *f*(*t*) are:
- (1) because it is grounded on generic SA and external properties, its sign and magnitude correlates with the true polarity of *t*, but
- (2) it is independent from the text patterns within *t* (conditional on *t*'s true polarity).

Estimating *P*(*Pol*(*w*) = A) - I

- Let there be a previously unknown text pattern *w*, appearing in $T_{SA\text{-}NEUTRAL}$. We are interested in probabilistically estimating the polarity of *w*, that is, in the value of *P*(*Pol*(*w*) = A), where A∈{POS, NEG}.
- Let $T_w = \{ t \in T_{SA\text{-}NEUTRAL} : w \in t \}$ be the set of all messages containing *w*. Then the probability *P*(*Pol*(*w*) = *A*) can be estimated from *f*(*t*) scores of messages in T_w:

$$P(Pol(w) = A \mid T_w, f) = \frac{P\left(f(T_w) \middle| Pol(w) = A\right)}{P(f(T_w))} * P(Pol(w) = A).$$

- The constant prior *P*(*Pol*(*w*) = *A*) can be ignored, assuming the set T_w is sufficiently large. In the main factor, we can safely assume that the different twits in T_w are independent from each other, so:

$$P(Pol(w) = A \mid T_w, f) \sim \prod_{t \in T_w} \frac{P\left(f(t) \middle| Pol(w) = A\right)}{P\left(f(t)\right)}.$$

Estimating *P*(*Pol*(*w*) = A) - II

- The marginal *P*(*f*(*t*)) can be estimated directly from the SVM classifier's scores on *T*. In the other part of the formula above, we are only dealing with twits that contain *w* whose polarity is non-neutral. According to our simplifying assumptions, we proceed as if there could be no conflicts, and the polarities of all twits in T_w are equal to the polarity of *w*. And then, the likelihood *P*(*f*(*t*) | *Pol*(*w*)=A) can be reduced to:

$$P(f(t) \mid Pol(w) = A) \approx \frac{P(f(t) \& Pol(t) = A)}{P(Pol(t) = A)}$$

- The constant marginal *P*(*Pol*(*t*) = *A*) can be directly estimated from a manually labeled development test set. And the rest can be estimated using the SA-polarized sets.

Estimating *P*(*Pol*(*w*) = A) - III

- Due to the conditional independence of *SA*(*t*) from *f*(*t*), given the true polarity of *t*, we have:

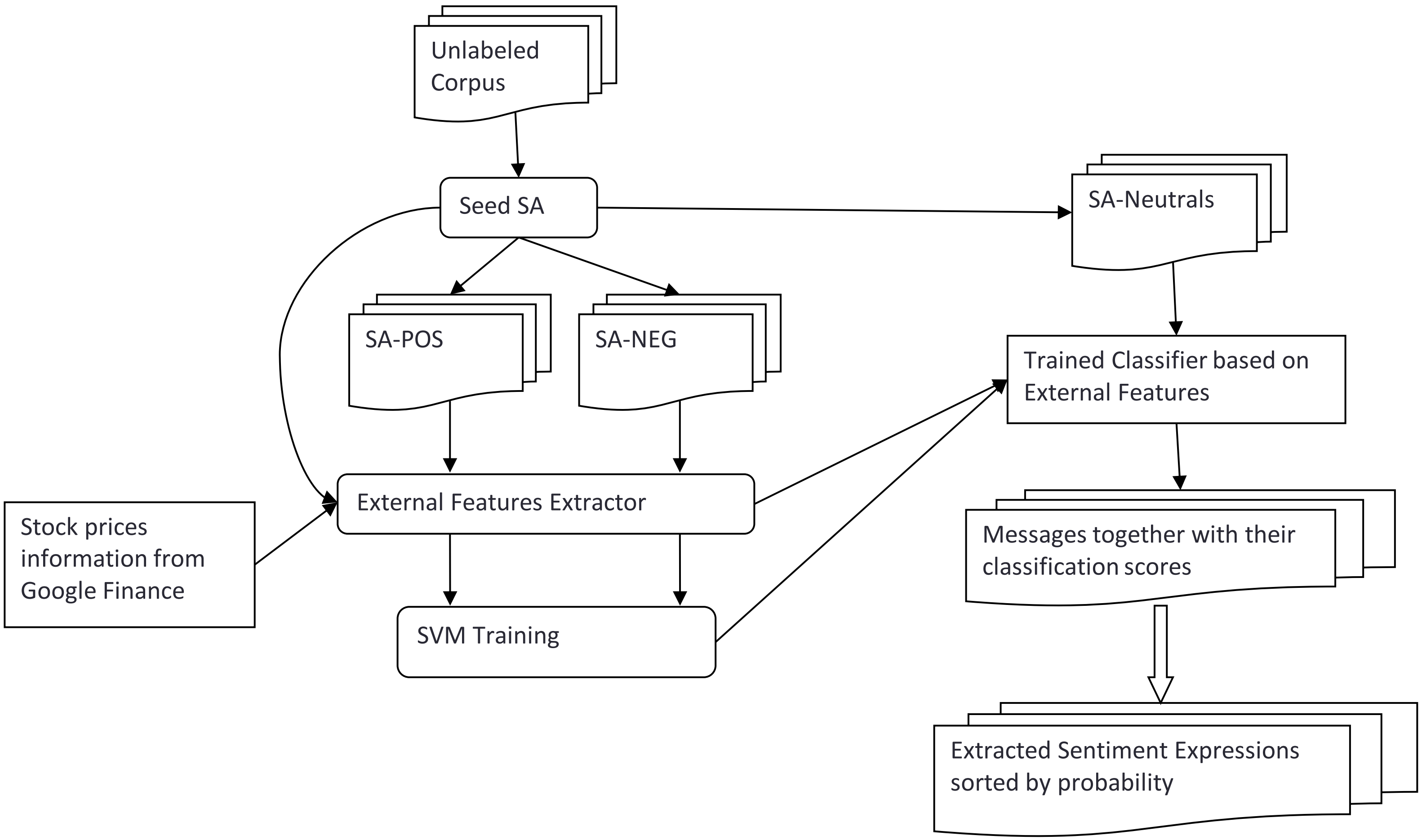
$$P(f(t) \& Pol(t) = A)$$

$$= \sum_{B \in POLS} P(f(t) \& Pol(t) = A \,\&\, SA(t) = B)$$

$$= \sum_{B \in POLS} P(f(t) \mid SA(t) = B) * P(SA(t) = B \mid Pol(t) = A) * \\ * \; P(Pol(t) = A)$$

- *P*(*f*(*t*) | *SA*(*t*)=*B*) and *P*(*SA*(*t*)=*B*) are estimated directly from the SA results and SVM results.
- *P*(*SA*(*t*) = *B* | *Pol*(*t*) = *A*) is estimated on a development test set.

architecture of the learning system


Unlabeled Corpus
Seed SA
SA-Neutrals
SA-POS
SA-NEG
Trained Classifier based on External Features
External Features Extractor
Stock prices information from Google Finance
Messages together with their classification scores
SVM Training
Extracted Sentiment Expressions sorted by probability

Features for the "External" View

- The external view can use any message properties that are independent from the message text.

1. **Stock-price-related features**. These are related to the price of the stock referenced in the messages, within some time frame of the message post time. The numerical values of the stock prices cannot be used directly, because they vary widely from stock to stock. However, we can identify and use 'price change events' – the points in time where the price of a stock significantly changed from one day to the next.
2. **Seed SA-related features.** These features more directly utilize connections between messages established by identity of their authors and/or subject.

Stock-price-related features

- There are many different possible adjustable parameters for identifying the useful price changes, and there is no a priori reason to select any particular numbers. Therefore, we choose several different reasonable values, and let an SVM classifier training algorithm choose the best. We use all possible combinations of the following:

1. SMALL is when the price is changed by 2-5%, LARGE is when the change is at least 5%, NOCHANGE if the change is less than 1%.
2. YESTERDAY is when the change occurs between yesterday's closing price (relative to the message post time) and today's opening price. TODAY is for changes between today's opening and closing prices, and TOMORROW is for changes between today's closing and tomorrow's opening prices.
3. PLUS when the price is increased, and MINUS when it decreased.

- We use all combinations of these properties as binary features, and also all possible intersections of pairs of them.

Seed SA-related features

- Let POS, NEG, NEUTRAL stand for the binary property of some message's (not the target's!) having overall positive, overall negative, and overall neutral sentiment labeling, respectively, according the seed SA. Also, let HASPOS, HASNEG, and HASANY stand for there being some positive, some negative, and some polarized sentiment expression within. (Thus, for example, POS and HASNEG may both be true for a message – if a negative sentiment expression occurs within, but the overall sentiment is positive).
- Given some set of messages, let EXIST_X be a binary property, of at least one of the messages satisfying X from (i). Also, let SUM_X, and AVERAGE_X be the real-valued sum and average of X over the messages in the set.

Seed SA-related features

- Given a target message, let DAYBEFORE, DAYAFTER, WITHINDAY, WITHINHOUR, WITHIN10MIN be the sets of messages posted the day before the message, the day after, within the same day, within the same hour, and within 10 minutes, respectively.
- Given a target message, let SAMEAUTHOR and ANYAUTHOR be the sets of messages posted by the same author and by any author, respectively.
- Again, we use all possible combinations of these properties as features.

Pattern-based Filtering

- This is a pattern-centric learning method, different from the message-based learning described above. The method uses a different learning model. It is based on the observation that different sentiment expressions occurring within the same message generally tend to be of the same polarity. Thus, given a candidate expression, we can ascertain its polarity by observing all messages that contain the expression which were labeled by the seed SA.
- The method cannot be directly incorporated into the above-described learning model, because it is directly using the message text, and does not satisfy the independence condition. However, the method can be used to perform an additional filtering step for the learned expressions, significantly improving the precision of the overall learning process.

Sentiment Analysis Systems

- The basic architecture of the SA system determines the kinds of sentiment expressions that can be used and learned. It is also the seed SA used for starting off the learning process and for calculating the SA-related features.
- Since the goal is seeking new polarity expressions, the SA system must be able to use them directly. This immediately eliminates from consideration some SA architectures, such as bag-of-words classification-based. (They should also be eliminated on independent grounds)
- Mainly, we experiment with a SA architecture, which is based on CARE-II-HPSG parser and relation extraction system.

Sentiment Analysis Systems

- We experiment with two versions of the system: GenericSA, a general-purpose SA system, which contains only language-wide sentiment expressions, and FinancialSA, which is an extension of GenericSA, created by manually adding many financial domain-specific sentiment expressions.
- For baseline, we also use DictionarySA, a simple system that classifies a message into positive, negative, or neutral categories according to the number of sentimental expressions that occur within the message. The expressions are the words and multi-word sequences taken from the GenericSA, without the additional syntactic information.

Care-II

- CARE-II is a domain-independent framework for building Information Extraction systems. The framework includes a grammar description language and the supporting tools.
- The core of the framework is a parser, which is capable of parsing arbitrary weighted typed-feature-structure context-free grammars (WTFSCFG-s).
- These are weighted CFG-s, in which every matched symbol, terminal or non-terminal, carries a typed feature structure; the grammar rules have access to the feature structures of their component symbols, building from them the feature structures for their heads, by applying the operations of unification, slot extraction, and slot removal.

Care-II HPSG

- CARE-II-HPSG is an English grammar written for the CARE-II framework, based on the principles of HPSG grammar theory.
- The grammar's lexicon is largely underspecified. Only the most frequent and functional words have full definitions, while the open classes of words are defined using generic underspecified lexical entries and tightly-integrated feature-rich sequence classification models for parts-of-speech (POS) and named entities recognition (NER).
- The models provide weights for different possible typed-feature-structure assignments. Then, for any input sentence, the parser generates a single highest-weight parse – the parse which is the most consistent with both the grammar rules and the NER and POS classifiers.

Using CARE-II-HPSG for Sentiment Analysis

- When used for sentiment analysis, either general purpose or domain specific, the lexicon is extended to include sentiment words and expressions, which include sentiment labels in the semantic parts of their HPSG feature structures. The labels may indicate polarity of expressions, their intensity, and their combining properties, such as behavior under negation.
- After a parse of a sentence is generated, it is post-processed by the SA post-processor, which merges sentiments from related expressions, performs coreference resolution, and attaches the sentiments to their targets, where appropriate. The post-processor is rule-based and deterministic.

To Parse or not to Parse?

Benefits in using full parsing for the SA task:

1. Precise identification of sentiment target in cases where several entities are available as possible targets.
2. Principled and uniform combining interdependent sentiment expressions and processing of negation.
3. Disambiguation for the cases where polarity of an expression depends on its syntactic role and/or part-of-speech (as in, for example, "fine" as a positive adjective vs. "fine" as a negative noun).
4. Principled and uniform way of defining multi-word sentiment expressions - using syntactic and semantic links instead of simple words proximity.

- The disadvantages of using full parsing are: slower processing speed, possible problems with bad grammar and typos, and generally low quality of parses, which may introduce SA errors instead of solving them. However, with a robust parser, these disadvantages should be minimized.

GenericSA, FinancialSA, and DictionarySA

- FinancialSA is the same as GenericSA, but extended by manually adding many domain-specific lexical entries (for the financial domain, investor sentiments in particular).
- DictionarySA is provided as a baseline. It is a very simple SA system that contains a dictionary of sentiment words and word sequences, and classifies a text by counting the number of occurrences of various polarity expressions within it. Whichever polarity occurs most frequently wins. If the number of occurrences is equal, the message is considered neutral. The initial dictionary is the same as used in GenericSA, without the additional syntactic information, such as parts-of-speech, valence, etc.
- The type of SA determines what kind of sentiment expressions is available for learning. The simple DictionarySA can only learn words and multi-word sequences. The parser-based SA systems are able to learn more complex patterns.

Pattern Types

1. Word patterns: individual and compound non-proper nouns, verbs, and adjectives.
2. Valence patterns: head word together with its Valentes, which can be noun phrases or preposition phrases complemented by noun phrases. The noun phrases are identified by their head noun, which can be either unrestricted, or restricted to a specific common noun, or restricted to the sentiment target entity type (Company name or stock symbol for the financial domain).
3. Modifier patterns: a head word modified by an adjectival phrase or a prepositional phrase complemented by a noun phrase. Same restrictions apply as for valence patterns.

Experimental Evaluation

- We use a corpus of several million stocks-related twitter messages collected between May and October 2011. We only use the messages related to stocks for which we were able to collect the price information from Google Finance.
- For the test set we use a manually-labeled set of randomly chosen 1500 twits. Another set of 500 twits was used as a development test set for estimating the marginal probabilities and for tuning the final threshold parameter.

Baseline

- In the baseline experiment we compare the results produced by the three seed SA systems that we use. For reference, we also show the results produced by Bing's system on the same test set. The system is one of the state-of-the-art general-purpose SA systems.
- Ding, X. et al. 2008. A Holistic Lexicon-Based Approach to Opinion Mining. *the Conference on Web Search and Web Data Mining (WSDM-2008)* (2008), 231–239.

Initial Results on StockTwits (Base Line)

	TP	FP	FN	Prec	Recall	F1
Bing's	365	295	418	0.553	0.466	0.506
Dictionary	308	223	474	0.580	0.393	0.468
GenericSA	159	94	506	0.628	0.239	0.346
FinancialSA	284	116	381	0.710	0.427	0.533

Analysis

- As can be seen from the table, generic SA systems have relatively low accuracy, due to specifics of the domain.
- Also notable, that the simple dictionary-based SA is not worse in precision from a much more sophisticated Bing's system in this domain (although much worse in recall).
- Note, that in the Table 1 we only consider positive and negative sentiments when counting the "true positives" (TP). Neutral sentiments are not included in the evaluation, except when they contribute to "false positives".

Results with neutral messages included

	TP	FP	FN	Prec	Recall	F1
Bing's	1239	295	418	0.808	0.748	0.777
Dictionary	1183	223	474	0.841	0.713	0.771
GenericSA	964	94	506	0.911	0.655	0.762
FinancialSA	1066	116	381	0.901	0.736	0.810

Learning Experiment

- In this experiment we compare the results produced by learning new sentimental expressions from the twitter messages corpus.
- In addition to comparing the learning capabilities of the three SA systems, we also compare three sets of external features:
- (1) PriceOnly set, which contains only features based on stock price,
- (2) SeedSA-Only set, which contains only features based on using Seed-SA-produced classification of messages related to the same entities, and
- (3) Full set, containing both Price-related and SeedSA-related features, as well as their intersections.
- The experiments are performed in the following way: given a SeedSA and an external feature set, we process the corpus with the SA, producing three separated sets of messages: SA-POS, SA-NEG, and SA-NEUTRAL. Then we train an SVM classifier using the representations in the external feature set of SA-POS and SA-NEG as training data. This classifier is applied on the representations of the messages in SA-NEUTRAL, producing a score for each message within.

Learning Experiment

- Then, for every candidate expression that appears at least ten times for different stocks (so as to eliminate local biases), we calculate its two final probability scores (of being positive and of being negative), using the formulas in Section 4. Given a probability threshold, we append all discovered expressions that pass the threshold to the corresponding seed system, and perform the test.
- We found that the final score number, while good for ordering the expressions, is not very indicative numerically. Consequently, we select the best thresholds using a development test set.
- Finally, we apply the extended SA system on the test set. The results are shown in Table 3.

Learning Experiment

- As can be seen from the table, improvements in accuracy are achieved for all of the seed SA-s using any of the feature sets. Somewhat surprisingly, FinancialSA shows the most improvement, even though it is the best of the systems from the beginning. This is probably due to the fact that it produces significantly better initial training sets.
- For GenericSA and FinancialSA, as expected, the Full feature set produces significantly better results than using either Price or SA-based feature sets separately. Unexpectedly, for DictionarySA, all three feature sets produce very similar results. The reason for the difference is unclear, and is under investigation.

Results of the Learning Experiment

	Dictionary			Generic SA			Financial SA		
	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Baseline	0.580	0.393	0.468	0.628	0.239	0.346	0.710	0.427	0.533
Price	0.599	0.427	0.498	0.635	0.252	0.361	0.697	0.443	0.541
SA-based	0.588	0.423	0.492	0.643	0.254	0.364	0.699	0.443	0.542
Full	0.589 (+0.9%)	0.428 (+3.5%)	0.495 (+2.7%)	0.635 (+0.7%)	0.262 (+2.3%)	0.370 (+2.4%)	0.700 (-1.0%)	0.518 (+9.1%)	0.595 (+6.2%)

Conclusions and Future Work

- We experimented with several SA systems and different external features to identify domain-specific expressions in financial messages on social media. We proposed a novel unsupervised multi-view-based approach that uses a seed SA system together with domain-specific external information in order to mine a large corpus of messages for domain-specific sentiment expressions, which, in turn, can extend the SA to improve its accuracy.
- We contribute to the body of knowledge on sentiment analysis, in general, and on sentiment analysis of financial social media messages, in particular. The proposed unsupervised methodological approach to sentimental analysis, which uses multiple views, may be adapted to other studies from other domains requiring the solving of sophisticated sentiment analysis problems.

Conclusions and Future Work

- Our experimental results indicate that our method is successful in integrating diverse sources of external information for the learning purposes. The sources we compare are stock prices on the one hand, and SA results on messages related by subject or by author, on the other. Our results show that when combined in our approach, the sources produce much better accuracy than individually, at least for the best-performing SA systems.
- Future studies may incorporate further external views' features, such as events known to have a positive or negative effect on companies, or may address both the time reference mentioned in the messages and the price movement in a corresponding period in order to generate the domain-specific lexicon.

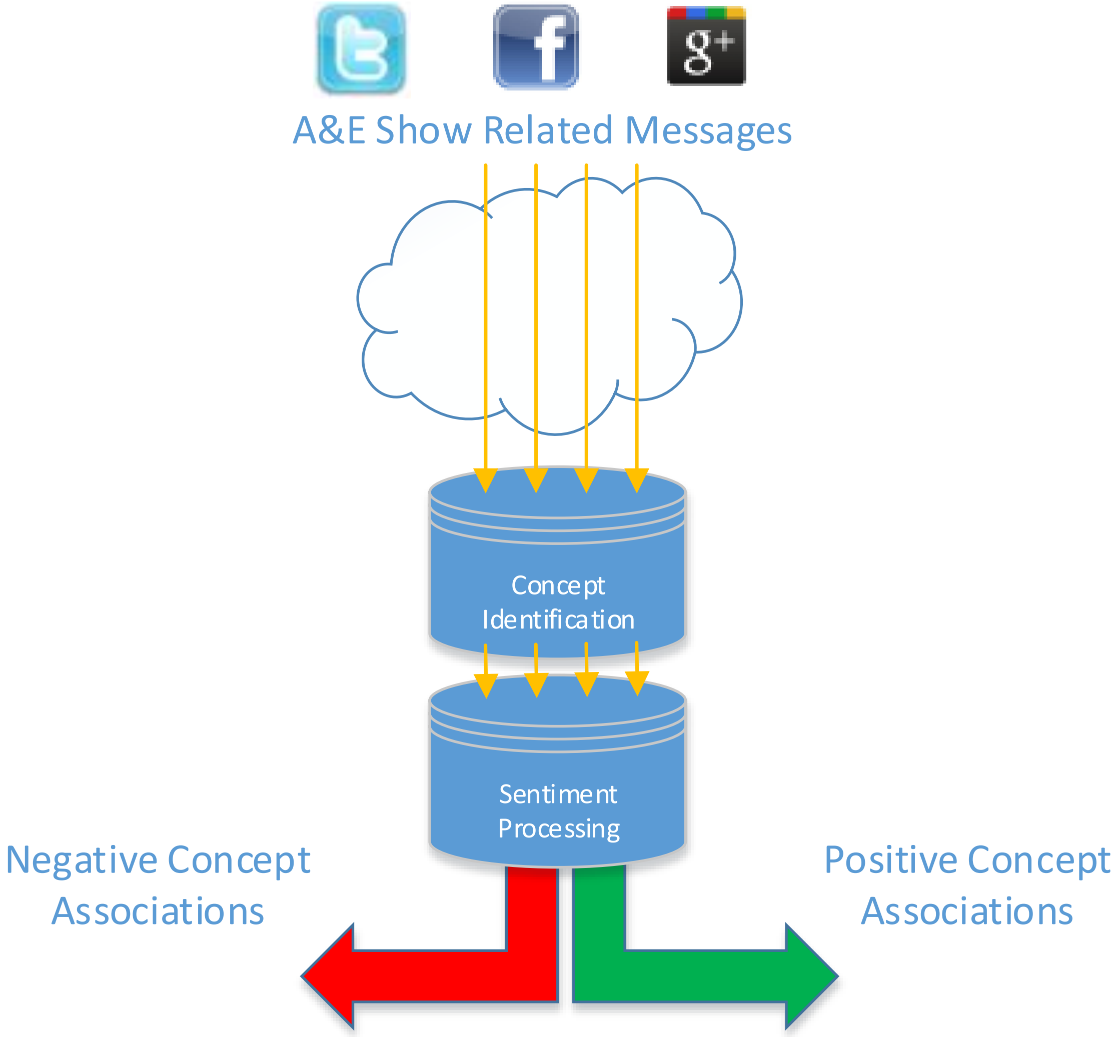
Content Profiling

Social Media Mining

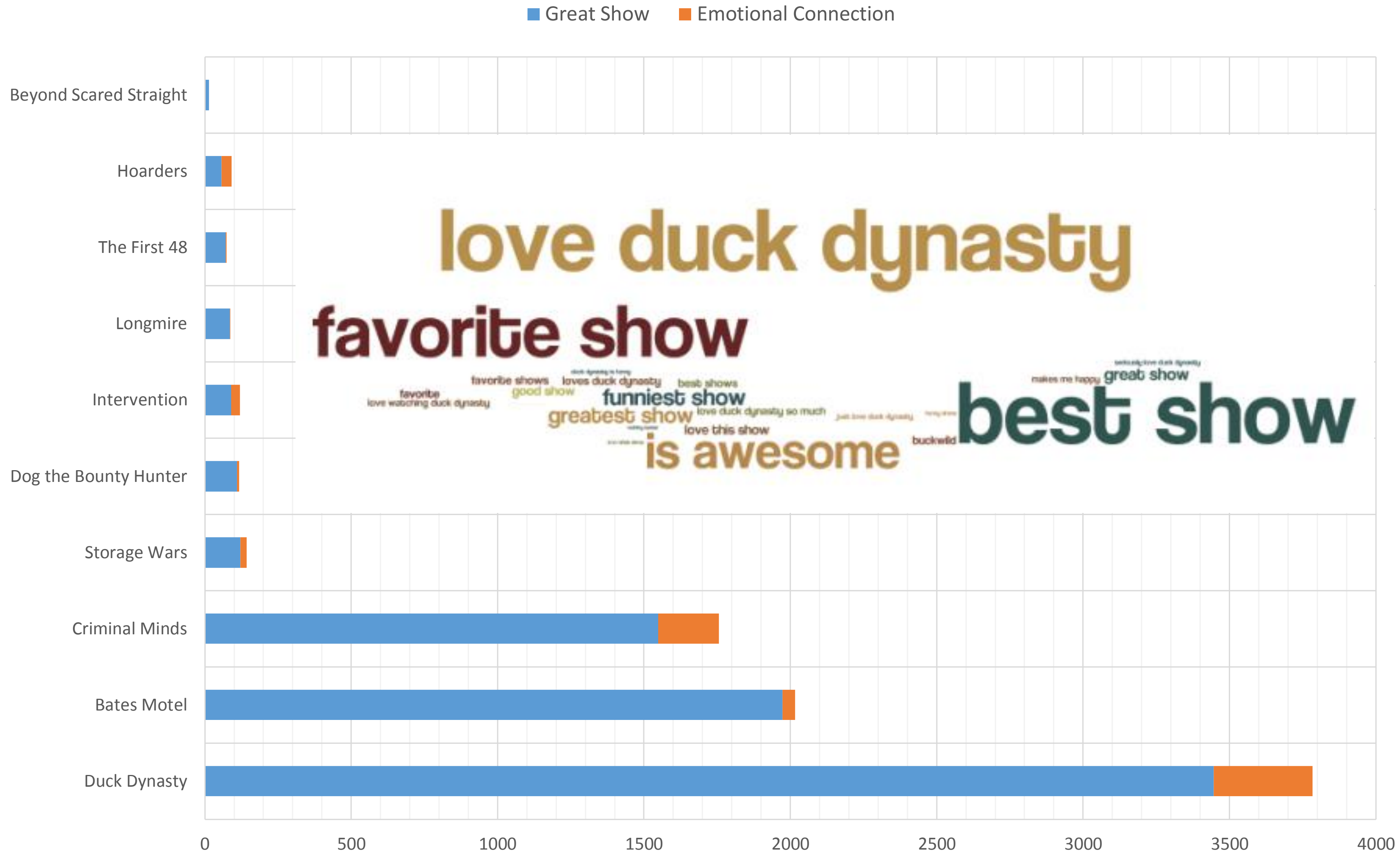
Concept Mining

- Understand the negative and positive concepts that consumers associate with top shows in their tweets, Facebook and Google+ updates
- Visualize and track the trending concepts associated with each show over time

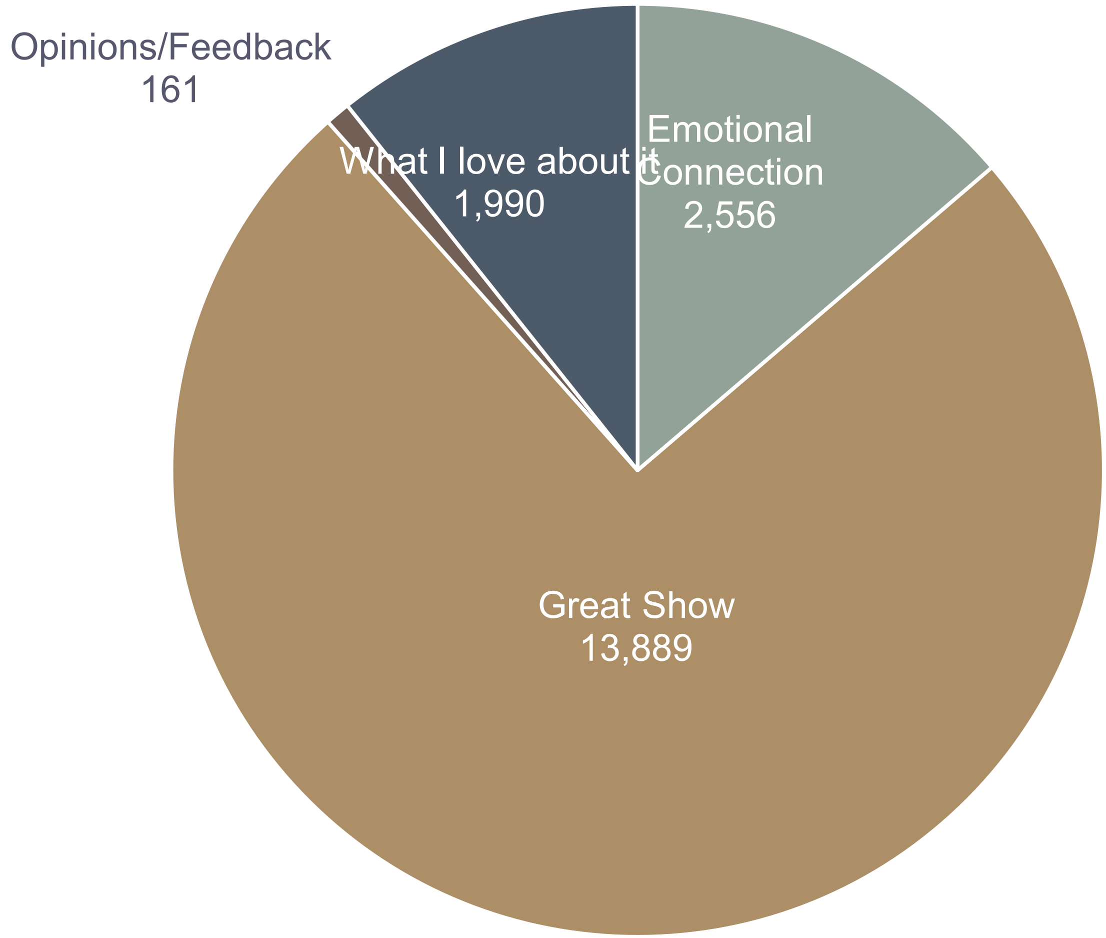
Concept Sentiment


A&E Show Related Messages
Concept Identification
Sentiment Processing
Negative Concept Associations
Positive Concept Associations

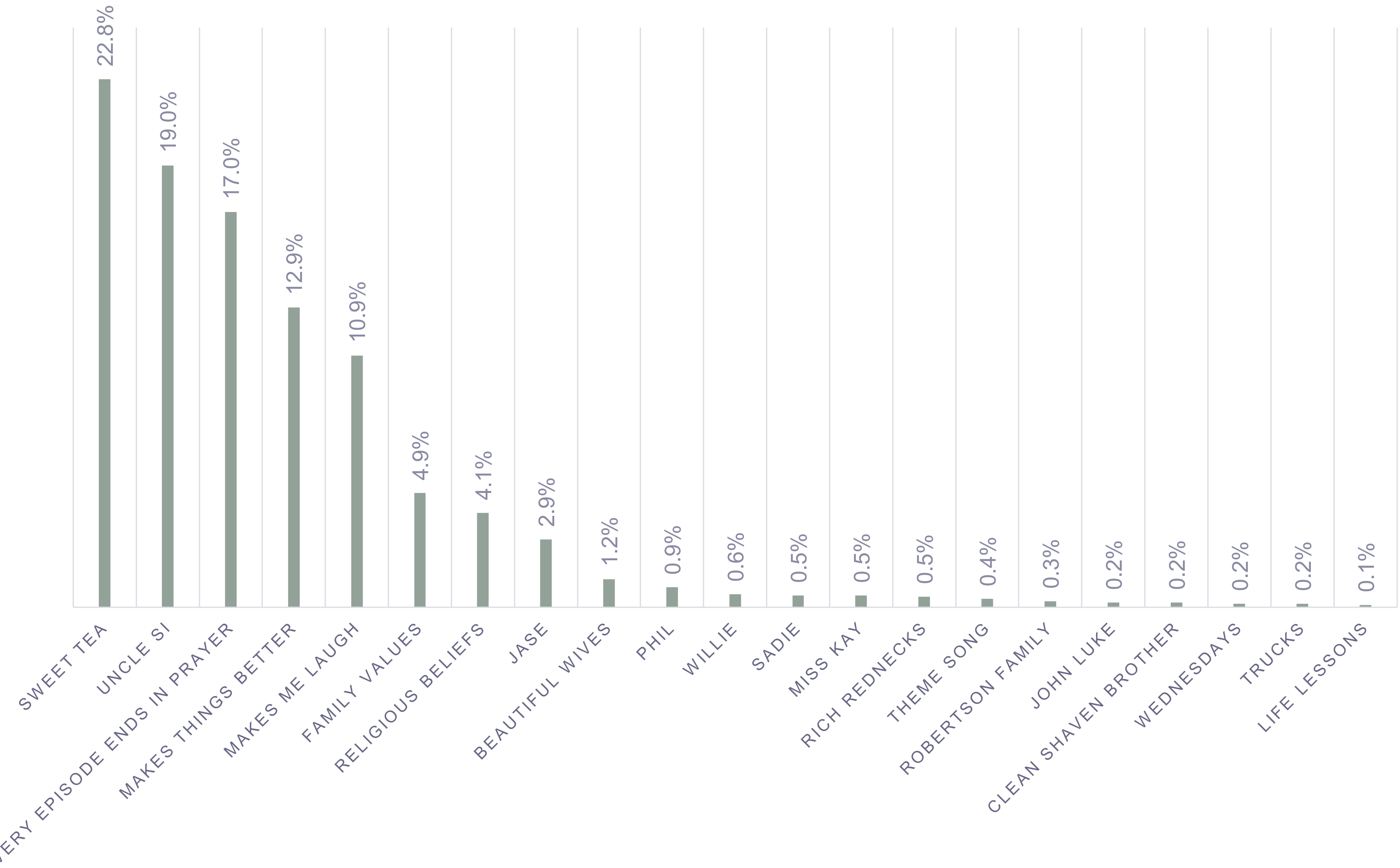
Positive Expression Categories – A&E Shows

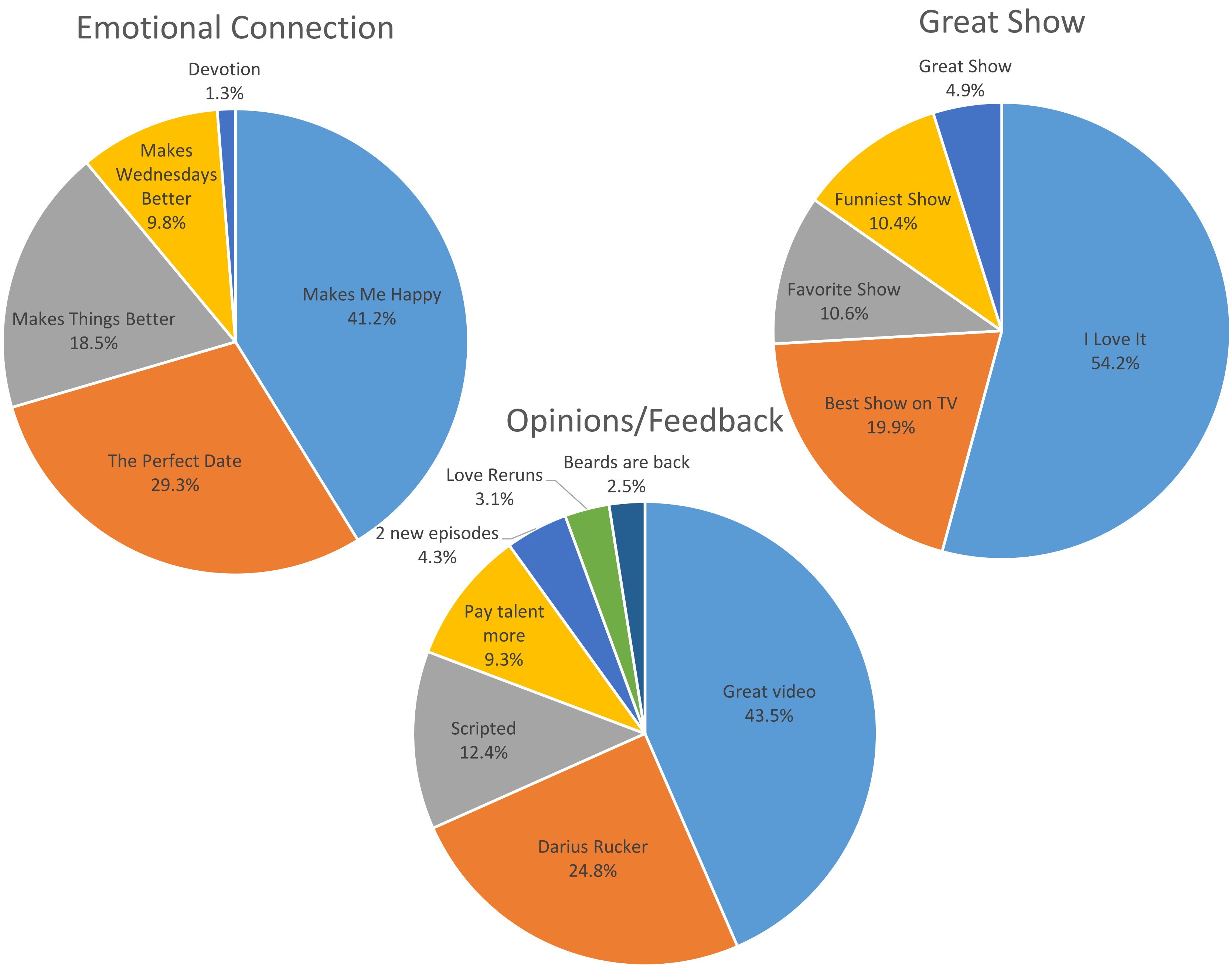


Positive Concept Categories



WHAT I LOVE ABOUT DUCK DYNASTY





Emotional Connection

Great Show

Opinions/Feedback

RESEARCH INSIGHTS

How We Can Impact Research

- Provide real time information regarding your target audience
- Identify issues before longer term research can be fielded and reported
- Opportunity to utilize social media chatter to establish the drivers of popularity—not just that there are conversations, but what's being talked about
- Provide overall chatter to inspire what comes next topics

Research Insights

- Social Media Popularity Tracking
  - Track daily by show the total mentions, positive and negative across Facebook, Google+ and Twitter
    - Understand each show's daily Pos/Neg ratio – the true measure of a program's resonance with your audience
- Crossover Profiling
  - Understand the off-network shows that consumers mention most frequently with your shows across social media
    - Understand both positive and negative frequencies

Social Media Mining

Popularity Tracking

The Social Media Popularity Ratio

The Popularity Ratio (Positive mentions/Negative mentions) is a much better indicator of a show’s popularity than buzz alone (total mentions).

Sorted	Mentions	Positives	Negatives	Neutral	Pos%	Neg%	Popularity
Shipping Wars	102	20	2	80	19.6%	2.0%	10.0
Southie Rules	69	32	6	31	46.4%	8.7%	5.3
Duck Dynasty	7972	1827	567	5578	22.9%	7.1%	3.2
Bates Motel	6578	1089	489	5000	16.6%	7.4%	2.2
Criminal Minds	4688	1063	492	3133	22.7%	10.5%	2.2
Storage Wars	1677	335	219	1123	20.0%	13.1%	1.5
Hoarders	1402	142	96	1164	10.1%	6.8%	1.5
Cold Case Files	38	8	6	24	21.1%	15.8%	1.3
Intervention	1775	285	214	1276	16.1%	12.1%	1.3
Barter Kings	72	15	12	45	20.8%	16.7%	1.3
Be the Boss	158	23	19	116	14.6%	12.0%	1.2
Beyond Scared Straight	470	57	56	357	12.1%	11.9%	1.0
American Hoggers	11	2	2	7	18.2%	18.2%	1.0
The First 48	651	80	96	475	12.3%	14.7%	0.8
Dog the Bounty Hunter	362	25	43	294	6.9%	11.9%	0.6
Longmire	156	26	62	68	16.7%	39.7%	0.4
Storage Wars: Texas	249	31	123	95	12.4%	49.4%	0.3
The Glades	45	1	11	33	2.2%	24.4%	0.1

Biggest Fans and Biggest…

Identify the authors who make the most comments about your shows: both positive and negative.

	Total	Positive	Negative	
Most Mentions (by single author)	98	24	94	
Author Stats				**Show**
Most Active Authors	BBallbags			Storage Wars: Texas
	einberg_kristi			Criminal Minds
	nellynichole			Criminal Minds
Most Negative Authors	BBallbags			Storage Wars: Texas
	BBallbags			Storage Wars
	fl0k_r0ck			Storage Wars: New York
Most Positive Authors	supremestream			Southie Rules
	Bobby6740			Storage Wars
	supremestream			Bates Motel