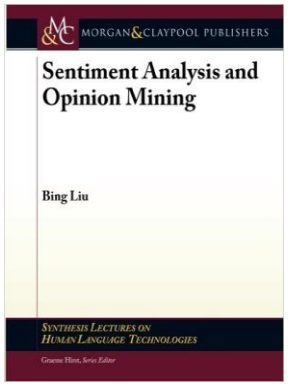
Principles/Social Media Mining

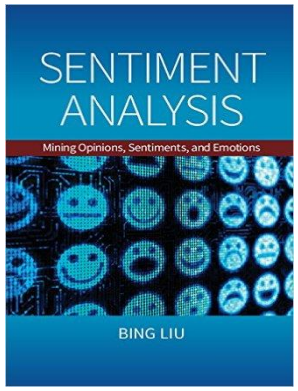
CIS 600

Week 9: Sentiment Analysis, Part 1: Data Mining Essentials & Overview

Edmund Yu, PhD
Associate Teaching Professor
esyu@syr.edu

October 20, 22, 2020



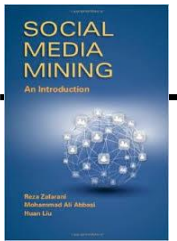


Data Mining

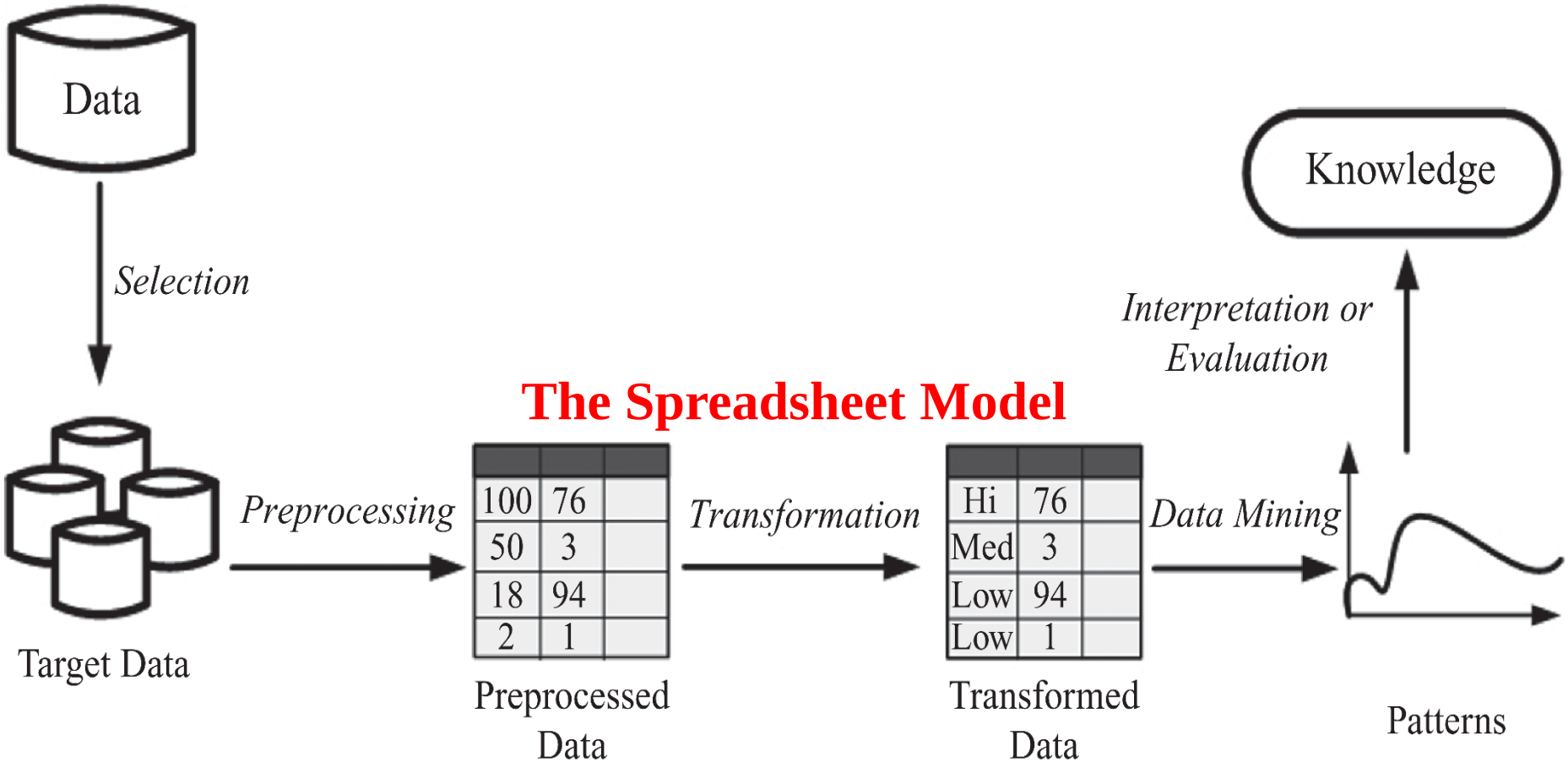
Data Mining: the **process** of discovering **hidden** and **actionable** patterns from data

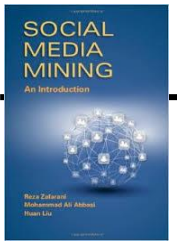
It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems

- Extracting/"mining" knowledge from large-scale data (big data)
- Data-driven discovery and modeling of hidden patterns in big data
- Extracting information/knowledge from data that is
  - implicit,
  - previously unknown,
  - unexpected, and
  - potentially useful



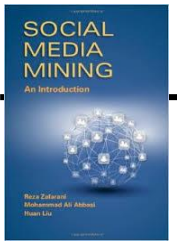
The KDD Process





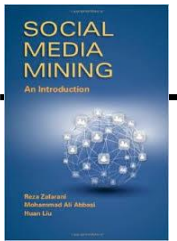
Data Mining vs Databases

- **Data mining** is the process of extracting hidden and actionable patterns from data
- **Database systems** store and manage data
  - Queries return part of stored data
  - Queries do not extract hidden patterns
- Examples of querying databases
  - Find all employees with income more than $250K
  - Find top spending customers in last month
  - Find all students from *engineering college* with GPA more than average



Examples of Data Mining Applications

- **Fraud/Spam Detections:** Identifying fraudulent transactions of a credit card or spam emails
  - You are given a user's purchase history and a new transaction, identify whether the transaction is fraud or not;
  - Determine whether a given email is spam or not
- **Frequent Patterns:** Extracting purchase patterns from existing records
  - beer ⇒ diapers (80%)
- **Forecasting:** Forecasting future sales and needs according to some given samples
- **Finding Like-Minded Individuals:** Extracting groups of like-minded people in a given network



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Twitter bot

From Wikipedia, the free encyclopedia
(Redirected from Twitterbot)

A **Twitter bot** is a type of bot software that controls a Twitter account via the Twitter API.[1] The bot software may autonomously perform actions such as tweeting, re-tweeting, liking, following, unfollowing, or direct messaging other accounts. The automation of Twitter accounts is governed by a set of automation rules that outline proper and improper uses of automation.[2] Proper usage includes broadcasting helpful information, automatically generating interesting or creative content, and automatically replying to users via direct message.[3][4][5] Improper usage includes circumventing API rate limits, violating user privacy, spamming,[6] and sockpuppeting.

Contents [hide]

Features [edit]

It is sometimes desirable to identify when a Twitter account is controlled by a bot.[7] In a 2012 paper,[1] Chu et al. propose the following criteria that indicate that an account may be a bot (they were designing an automated system):

- "Periodic and regular timing" of tweets;
- Whether the tweet content contains known spam; and

Data

Data Instance

- In the KDD process,
  - Data is in a tabular format (a set of **instances**) → The Spreadsheet Model
- Each instance is a collection of attributes/features related to an object or person
  - A patient's medical record
  - A user's profile
  - A gene's information
- Instances are also called *points*, *data points*, or *observations*

Data Instance:

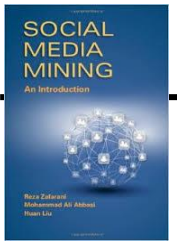
Attributes				Class
Name	Money Spent	Bought Similar	Visits	Will Buy
Mary	High	Yes	Rarely	Yes

Feature Value

Class Attribute

Features (Attributes or measurements)

Class *Label*



Data Instance

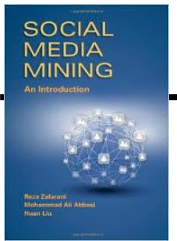
- Predicting whether an individual who visits an online book seller is going to buy a specific book

Attributes				Class
Name	Money Spent	Bought Similar	Visits	Will Buy
John	High	Yes	Frequently	?
Mary	High	Yes	Rarely	Yes

Unlabeled Example

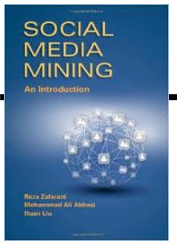
Labeled Example

- Features can be
  - **Continues**: values are numeric values
    - Money spent: $25
  - **Discrete**: can take a number of values
    - Money spent: {high, normal, low} → **Categorical**



Data Types

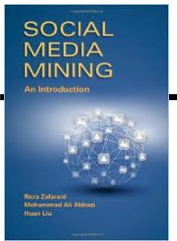
- **Nominal (Categorical)**
  - **Operations**:
    - Mode (most common feature value), Equality Comparison
  - E.g., {male, female}
- **Ordinal (Categorical)**
  - Feature values have an intrinsic order to them, but the difference is not defined
  - **Operations**:
    - same as nominal, feature value rank
  - E.g., {Low, medium, high}
- **Interval (Numerical)**
  - **Operations:**
    - Addition and subtractions are allowed whereas divisions and multiplications are not
  - E.g., 3:08 PM, calendar dates
- **Ratio (Numerical)**
  - **Operations:**
    - divisions and multiplications are allowed
  - E.g., Height, weight, money



Sample Dataset – Twitter Users

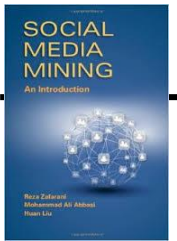
Activity	*Date Joined*	*Number of Followers*	*Verified Account?*	*Has Profile Picture?*
High	2015	50	FALSE	no
High	2013	300	TRUE	no
Average	2011	860000	FALSE	yes
Low	2012	96	FALSE	yes
High	2008	8,000	FALSE	yes
Average	2009	5	TRUE	no
Very High	2010	650,000	TRUE	yes
Low	2010	95	FALSE	no
Average	2011	70	FALSE	yes
Very High	2013	80,000	FALSE	yes
Low	2014	70	TRUE	yes
Average	2013	900	TRUE	yes
High	2011	7500	FALSE	yes
Low	2010	910	TRUE	no

Ordinal **Interval** **Ratio** **Nominal** **Nominal**



Text Data Representation

- The most common way to represent documents is to transform them into vectors
  - Process them with linear algebraic operations
- This representation is called "***Bag of Words***"
  - Vector Space Model
- Weights for words can be assigned by **TF-IDF**



Vector Space Model

- Consider a set of documents D
- Each document is a set (bag) of words
- **Goal**: convert these documents to vectors

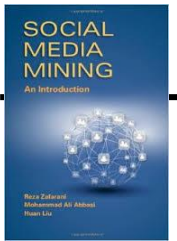
$$d_i = (w_{1,i}, w_{2,i}, \ldots, w_{N,i})$$

d_i : document i

$w_{j,i}$: the weight for word j in document i

How to set $w_{j,i}$

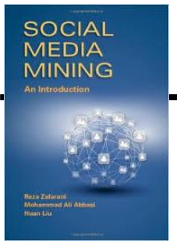
- Set $w_{j,i}$ to 1 when the word j exists in document i and 0 when it does not.
- We can also set $w_{j,i}$ to the number, of times the word j is observed in document i (frequency)



Vector Space Model: An Example

- **Documents:**
  - d_1: social media mining
  - d_2: social media data
  - d_3: financial market data
- Reference vector (**Dictionary**):
  - (social, media, mining, data, financial, market)
- Vector representation:

	social	media	mining	data	financial	market
d_1	1	1	1	0	0	0
d_2	1	1	0	1	0	0
d_3	0	0	0	1	1	1



TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF of term (word) t, document d, and document corpus D is calculated as follows:

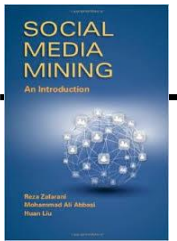
$$w_{j,i} = tf_{j,i} \times idf_j$$

$tf_{j,i}$ is the frequency of word j in document i

The total number of documents in the corpus

$$idf_j = \log_2 \frac{|D|}{|\{\text{document} \in D \mid j \in \text{document}\}|}$$

The number of documents where the term j appears



TF-IDF: A Simple Example

Document d_1 contains 100 words

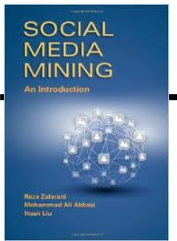
- Word "apple" appears 10 times in d_1
- Word "orange" appears 20 times in d_1

We have $|D| = 20$ documents

- Word "apple" only appears in document d_1
- Word "orange" appears in all 20 documents

$$tf - idf(\text{“}apple\text{”}, d_1) = 10 \times \log_2 \frac{20}{1} = 43.22$$

$$tf - idf(\text{“}orange\text{”}, d_1) = 20 \times \log_2 \frac{20}{20} = 0$$



TF-IDF: Another Example

Documents:

- d_1: social media mining
- d_2: social media data
- d_3: financial market data

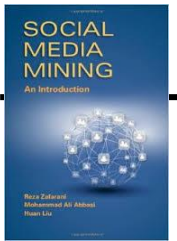
$$idf_{\text{social}} = \log_2(3/2) = 0.584$$
$$idf_{\text{media}} = \log_2(3/2) = 0.584$$
$$idf_{\text{mining}} = \log_2(3/1) = 1.584$$
$$idf_{\text{data}} = \log_2(3/2) = 0.584$$
$$idf_{\text{financial}} = \log_2(3/1) = 1.584$$
$$idf_{\text{market}} = \log_2(3/1) = 1.584$$

TF values:

	social	media	mining	data	financial	market
d_1	1	1	1	0	0	0
d_2	1	1	0	1	0	0
d_3	0	0	0	1	1	1

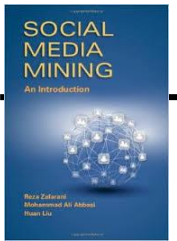
TF-IDF

	social	media	mining	data	financial	market
d_1	0.584	0.584	1.584	0	0	0
d_2	0.584	0.584	0	0.584	0	0
d_3	0	0	0	0.584	1.584	1.584



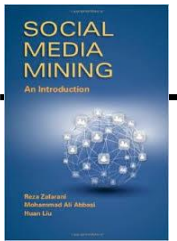
Data Quality

- When making data ready for mining, data quality needs to be assured
  - **Noise:** Noise is the distortion of the data
  - **Outliers**
    - Outliers are data points that are considerably different from other data points in the dataset
  - **Missing Values**
    - Missing feature values in data instances
    - **Solution:**
      - Remove instances that have missing values
      - Estimate missing values, and
      - Ignore missing values when running data mining algorithm
  - **Duplicate data**



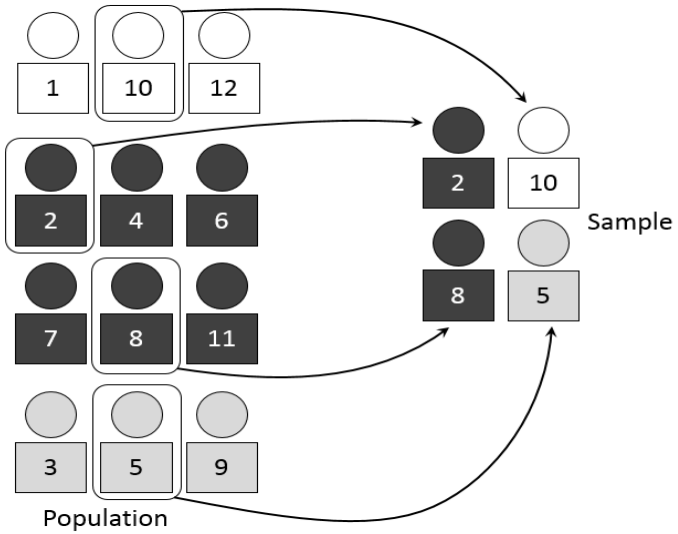
Data Preprocessing

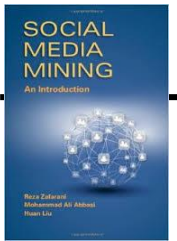
- **Aggregation**
  - It is performed when multiple features need to be combined into a single one or when the scale of the features change
  - Example: image width , image height -> image area (width x height)
- **Discretization**
  - From continues values to discrete values
  - Example: money spent -> {low, normal, high}



Data Preprocessing

- **Feature Selection**
  - Choose relevant features
- **Feature Extraction**
  - Creating new features from original features
  - Often, more complicated than aggregation
- **Sampling**
  - Random Sampling, with/without replacement
  - Stratified Sampling: useful when having class imbalance
  - Social Network Sampling





Data Preprocessing

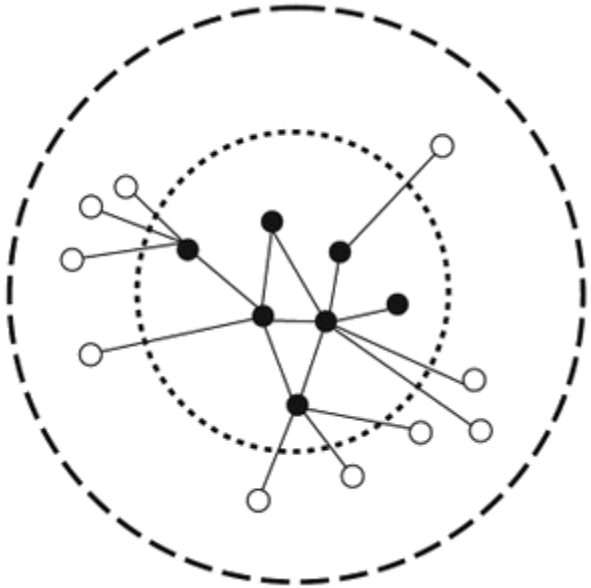
Sampling social networks:

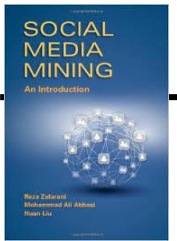
- Start with a small set of nodes (seed nodes)
- Sample

  (**a**) the connected components they belong to;

  (**b**) the set of nodes (and edges) connected to them directly; or

  (**c**) the set of nodes and edges that are within n-hop distance from them.





Data Mining Algorithms

- **Supervised Machine Learning Algorithms**
  - **Classification (class attribute is discrete)**
    - Assign data into predefined classes
      - Text categorization, sentiment analysis, etc.
  - **Regression (class attribute takes real values)**
    - Predict a real value for a given data instance
      - Predict the price of a given stock
- **Unsupervised Machine Learning Algorithms**
  - Group similar items together into some clusters
    - Detect communities in a given social network

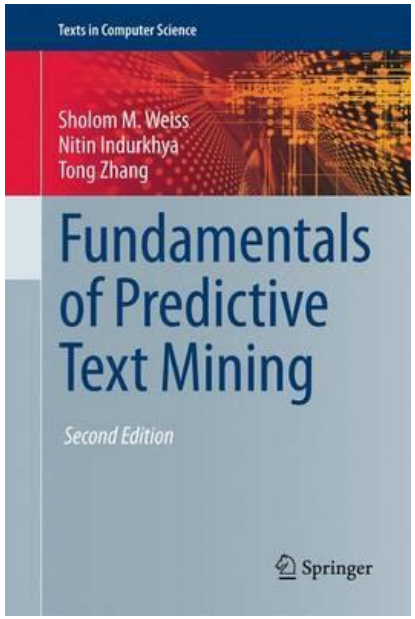
Note : Machine learning algorithms will be discussed separately later

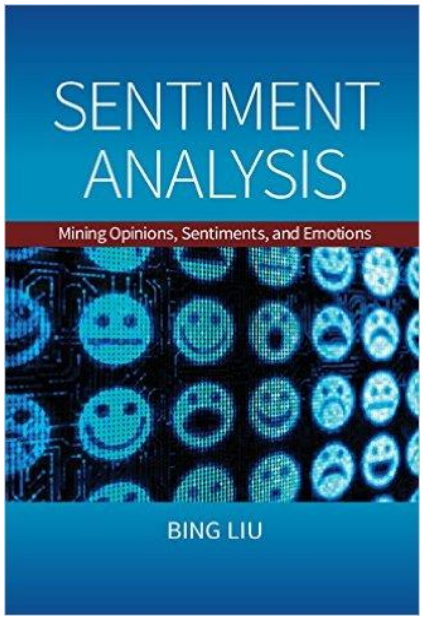
Reading Assignments

- Textbook #2 : Social Media Mining: An Introduction. By Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu, 2015, Cambridge. (Click its link in Blackboard.) **Chapter 5**.

Text Mining

`


Texts in Computer Science
Sholom M. Weiss
Nitin Indurkhya
Tong Zhang
Fundamentals of Predictive Text Mining
Second Edition
Springer


SENTIMENT ANALYSIS
Mining Opinions, Sentiments, and Emotions
BING LIU

Text Mining

- **Text mining**, also referred to as **text data mining**, roughly equivalent to **text analytics**, refers to the process of deriving high-quality information from **text**.
- High-quality information is typically derived through the devising of patterns and trends through means such as **statistical pattern learning**.

Text mining - Wikipedia, the free encyclopedia

*https://en.wikipedia.org/wiki/**Text_mining***

Text Mining vs Data Mining

- The text is usually a collection of **unstructured** documents with no special requirements for composing the documents. (Free text)

- Data-mining applications, on the other hand, deal with **structured** information.
  - The data must be prepared in a very special way before any learning methods can be applied. (See next slide)

Unstructured Data to Structured Data

- Most data-mining methods have proved remarkably successful without understanding specific properties of text such as the concepts of grammar or the meaning of words.
  - Only low-level frequency information is required, such as the number of times a word appears in a document.
- Hence, one of the main themes supporting text mining is the transformation of **text** into **numerical data**
  - The **unstructured** data become **structured**.

TF-IDF (Revisited)

Documents:

- d_1: social media mining
- d_2: social media data
- d_3: financial market data

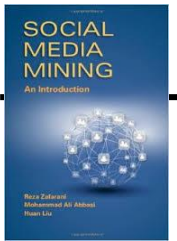
$$idf_{\text{social}} = \log_2(3/2) = 0.584$$
$$idf_{\text{media}} = \log_2(3/2) = 0.584$$
$$idf_{\text{mining}} = \log_2(3/1) = 1.584$$
$$idf_{\text{data}} = \log_2(3/2) = 0.584$$
$$idf_{\text{financial}} = \log_2(3/1) = 1.584$$
$$idf_{\text{market}} = \log_2(3/1) = 1.584$$

TF values:

	social	media	mining	data	financial	market
d_1	1	1	1	0	0	0
d_2	1	1	0	1	0	0
d_3	0	0	0	1	1	1

TF-IDF

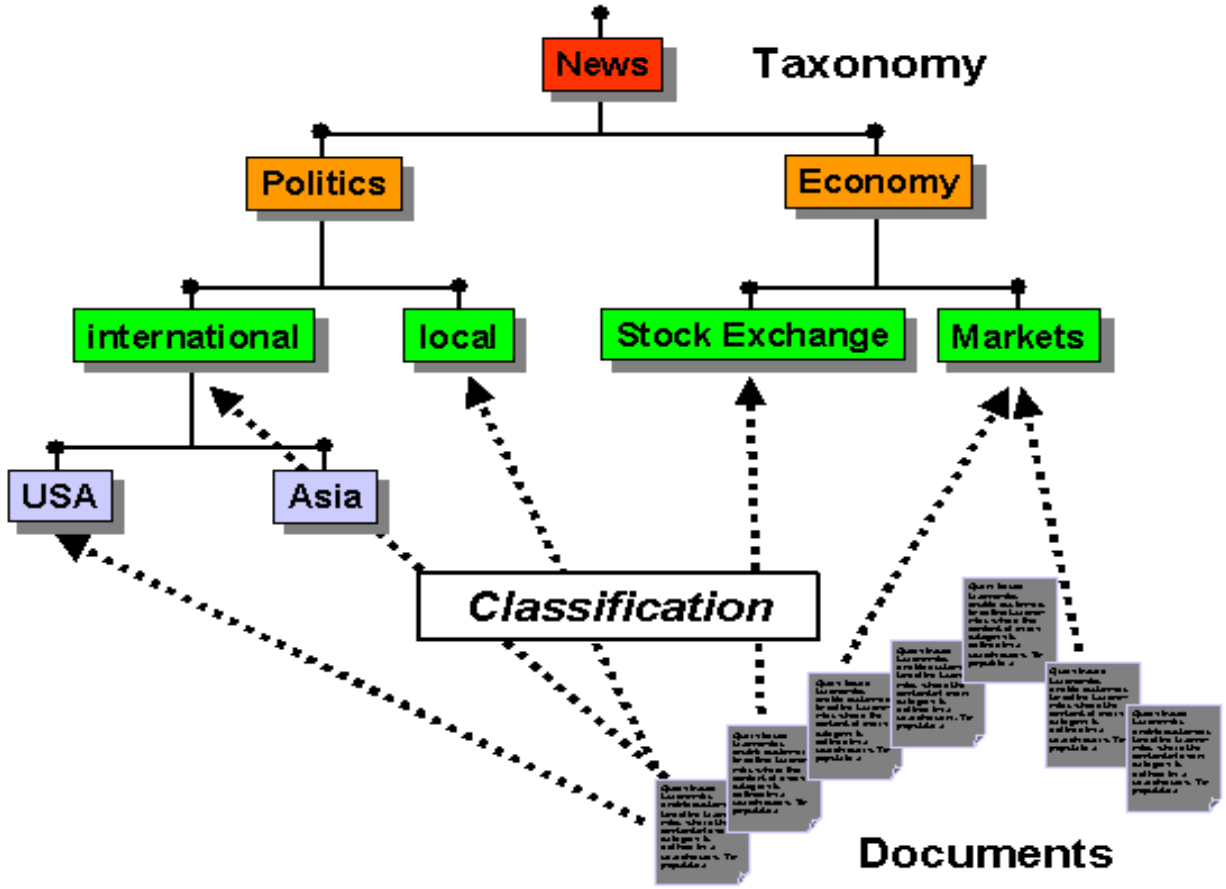
	social	media	mining	data	financial	market
d_1	0.584	0.584	1.584	0	0	0
d_2	0.584	0.584	0	0.584	0	0
d_3	0	0	0	0.584	1.584	1.584



Application Areas: Classification/Prediction

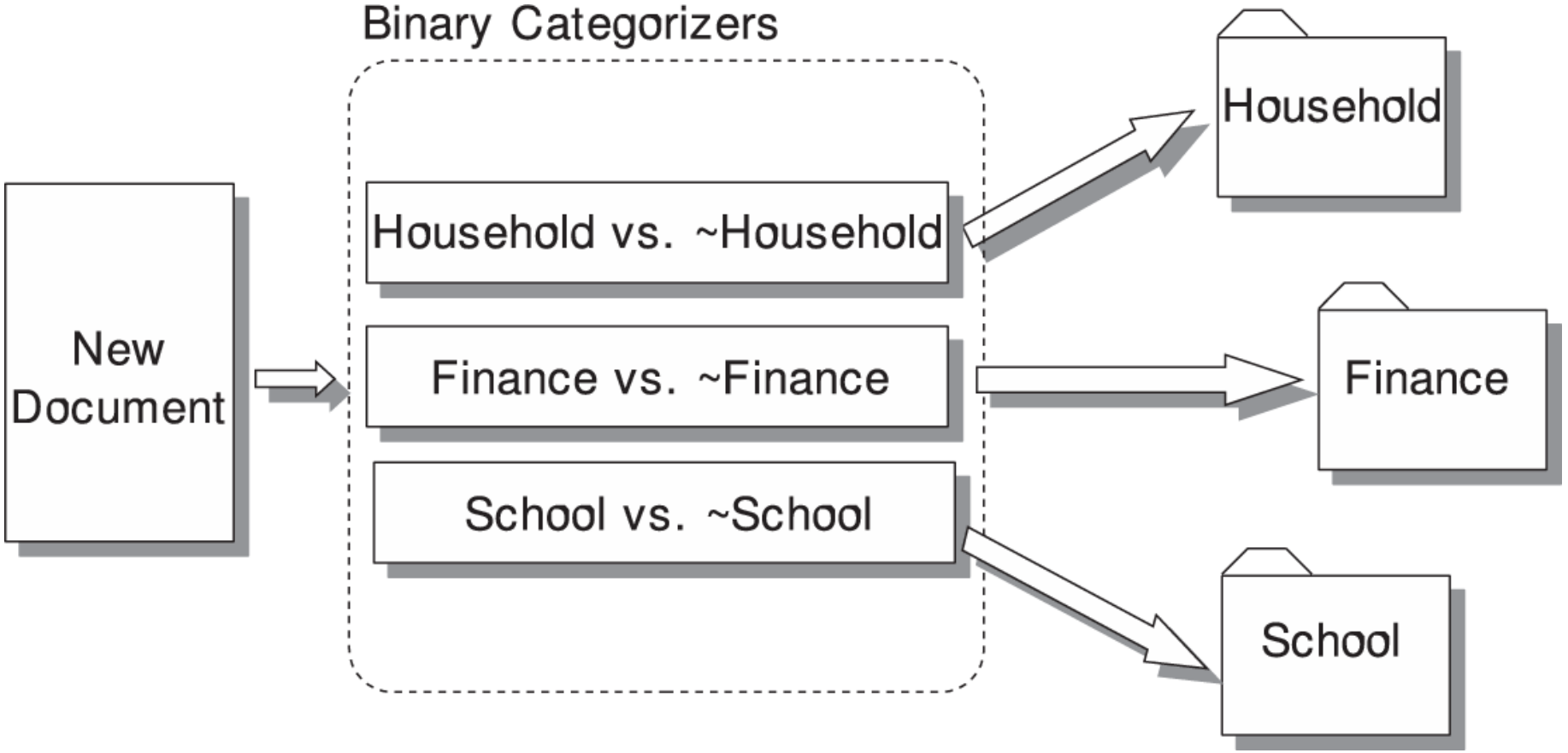
- Classification and prediction are among the most widely studied and applied methods and applications of text/data mining.
  - Given a sample of past examples and correct answers for each example, the objective is to find the correct answers for new examples. → **Text Categorization**
  - The concept of classification can be extended to data that do not have clearly labeled answers. → **Text Clustering**
    - Organize the data in such a way that we can make up labels or answers and expect these to hold in the future.
    - Although **similarity** between documents is an essential ingredient in organizing unlabeled documents into distinct groups, measuring similarity between documents is fundamental to most forms of document analysis, especially **Information Retrieval**.

Text Categorization

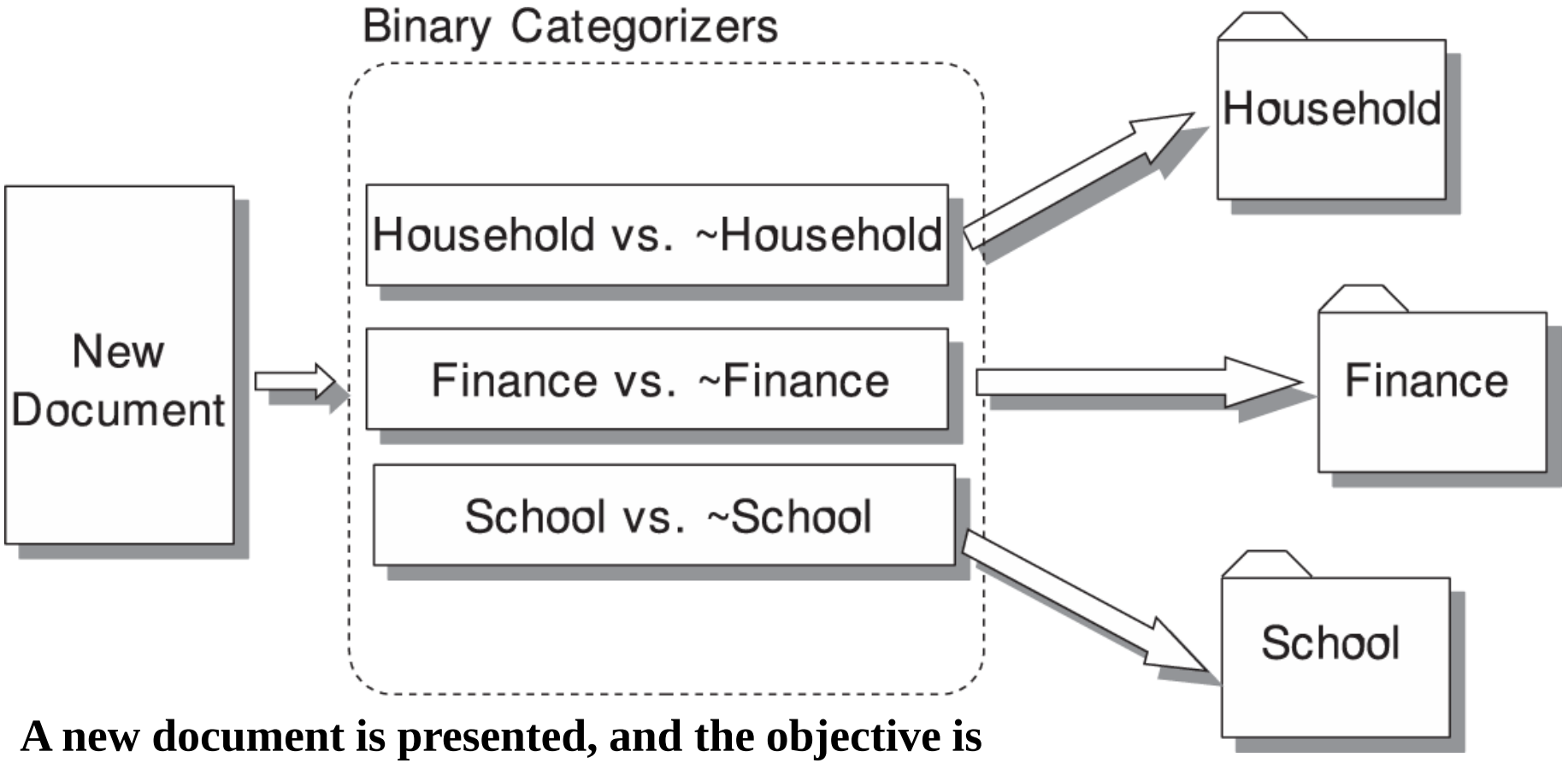


Text Categorization

Documents are organized into folders, one folder for each topic.



Text Categorization



A new document is presented, and the objective is to place this document in the appropriate folder.

Application Areas: Classification/Prediction

- Originally, text categorization was considered a form of indexing, much like the index of a book.
  - Nowadays, its applicability is much broader.
    - Automatically forwarding e-mail to the appropriate company department
    - Detecting spam mail, etc.
    - Predicting future stock movements based on news
      - You collect news articles that appeared prior to a rise or fall in stock prices along with company financial data. The labels would be binary, 1 for up and 0 for down
    - Or based on sentiments expressed in social media → **Sentiment Analysis**

Sentiment Analysis: An Example

http://www.sentiment140.com/search?query=iPhone+6s&hl=en

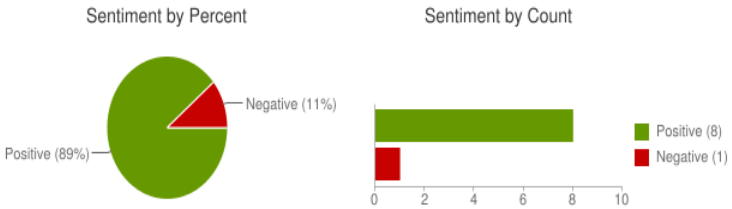
Sentiment140 - A Twitter Se...

Sentiment140

Tweet | Like 815 | G+1 202

iPhone 6s | English | Search

Sentiment analysis for iPhone 6s



Tweets about: iPhone 6s

johnnywhygull: I've entered to win an **iPhone 6S** 16GB Enter Here: https://t.co/HLab39lfK7
Posted: 1 minute ago

Galaxygamerone: RT @LeonKFox: I may as well stop denying it, I now know that I want an **iPhone 6S+** because I love iOS now. Likely going to wait for the 7 to?
Posted: 2 minutes ago

nonamejanee: What was I thinking this **iPhone 6s** Plus is way to big ?? can't even hold it with one hand
Posted: 3 minutes ago

BovsikT: RT @crazyboy1974: That's his **iPhone 6S**....!!! https://t.co/A2gTpsE4Oy
Posted: 3 minutes ago

UrduTweep: @ItnaSarah chota bhai hai, **iPhone 6s** us se ziyada thori hai, get him another one, sacrifice like 3 pairs of shoes >.>
Posted: 8 minutes ago

LoveChanyeol61: @Dalcomsoft_zone I want **iPhone 6s** ha ha??
Posted: 8 minutes ago

PhoneForSale: RT @tradeguide24: UNLOCKED BRAND NEW APPLE **IPHONE 6S** 128GB FOR SALE https://t.co/d6ijc5A7co #apple **#iPh**
Posted: 8 minutes ago

The results for this query are: Accurate | Inaccurate

Sentiment140

Search this site

API

We provide APIs for classifying tweets. This allows you to integrate our sentiment analysis classifier into your site or product.

Contents

Registration

You may register your application here: http://help.sentiment140.com/api/registration.

Please provide an `appid` parameter in your API requests. The `appid` value should be an email address we can contact. For example, if you are using the JSON Bulk Classification API, the HTTP endpoint that you use to send requests may look like this:

```
http://www.sentiment140.com/api/bulkClassifyJson?appid=bob@apple.com
```

where `bob@apple.com` is the main contact. You may URL-encode the `appid` value.

Technically you can use the API without supplying the `appid` parameter. But, we may block the requests that don't have an `appid` specified if we suspect abuse.

Commercial License

Deeply Moving: Deep Learning for Sentiment Analysis

This website provides a live demo for predicting the sentiment of movie reviews. Most sentiment prediction systems work just by looking at words in isolation, giving positive points for positive words and negative points for negative words and then summing up these points. That way, the order of words is ignored and important information is lost. In constrast, our new deep learning model actually builds up a representation of whole sentences based on the sentence structure. It computes the sentiment based on how words compose the meaning of longer phrases. This way, the model is not as easily fooled as previous models. For example, our model learned that funny and witty are positive but the following sentence is still negative overall:

This movie was actually neither that funny, nor super witty.

The underlying technology of this demo is based on a new type of Recursive Neural Network that builds on top of grammatical structures. You can also browse the Stanford Sentiment Treebank, the dataset on which this model was trained. The model and dataset are described in an upcoming EMNLP paper. Of course, no model is perfect. You can help the model learn even more by labeling sentences we think would help the model or those you try in the live demo.

Paper Title and Abstract

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank
Semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of composition. To remedy this, we introduce a

Paper: Download pdf

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)

Dataset Downloads:

Main zip file with readme (6mb)
Dataset raw counts (5mb)
Train,Dev,Test Splits in PTB Tree Format

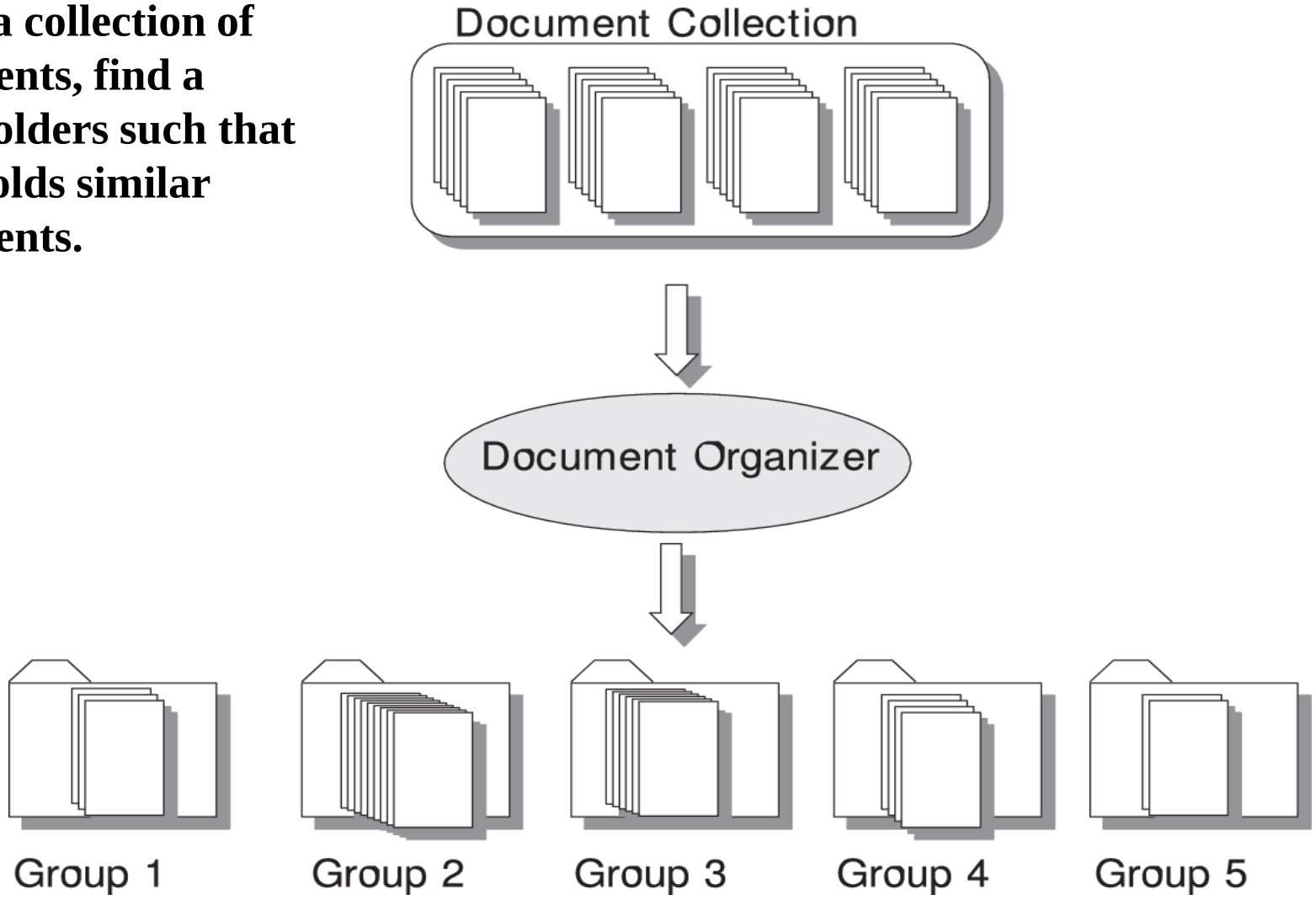
Code: Download Page

Press: Stanford Press Release

Dataset visualization and web design by Jason Chuang. Live demo by Jean Wu, Richard Socher, Rukmani Ravisundaram and Tayyab Tariq.

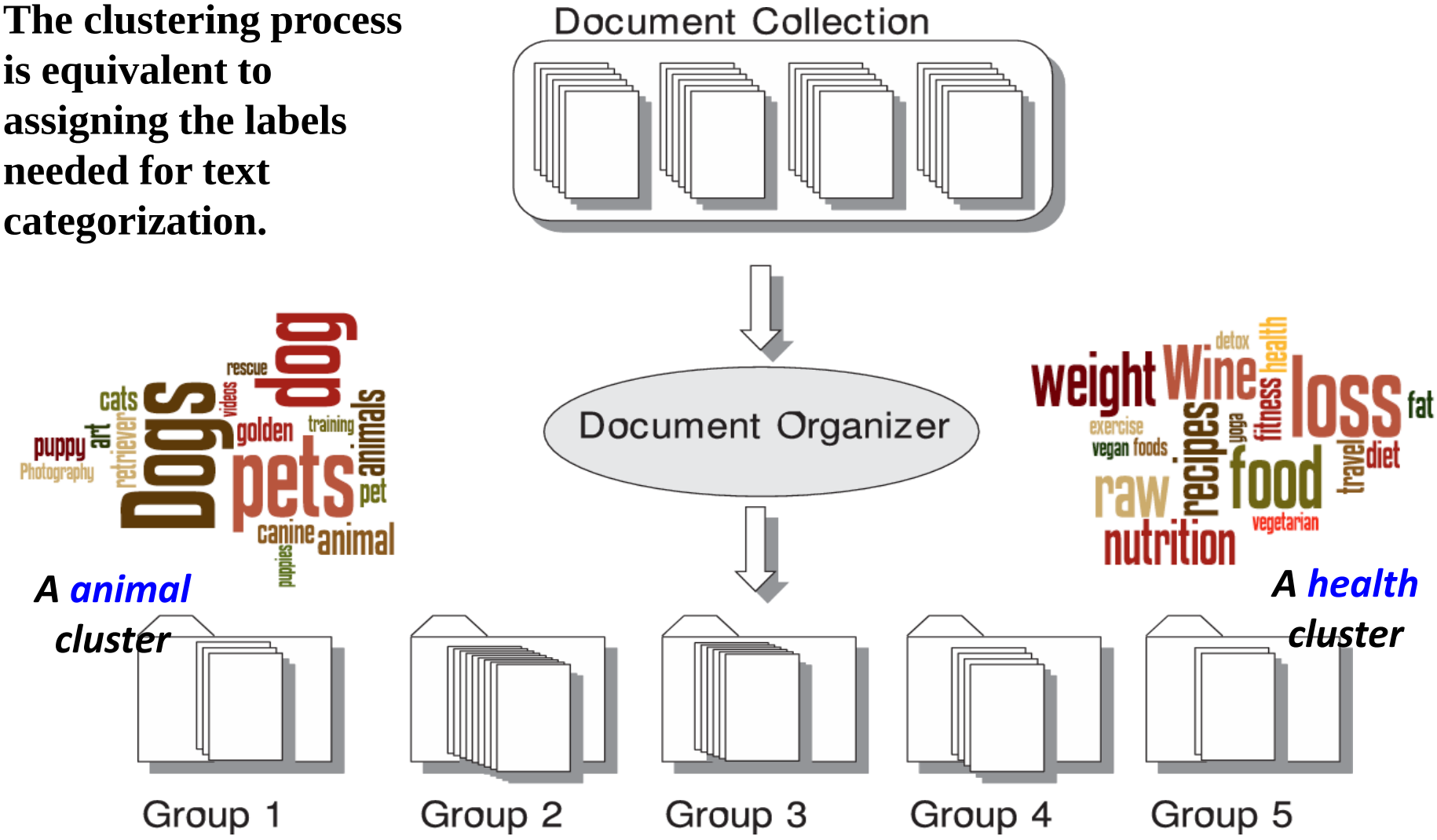
Text Clustering

Given a collection of documents, find a set of folders such that each holds similar documents.



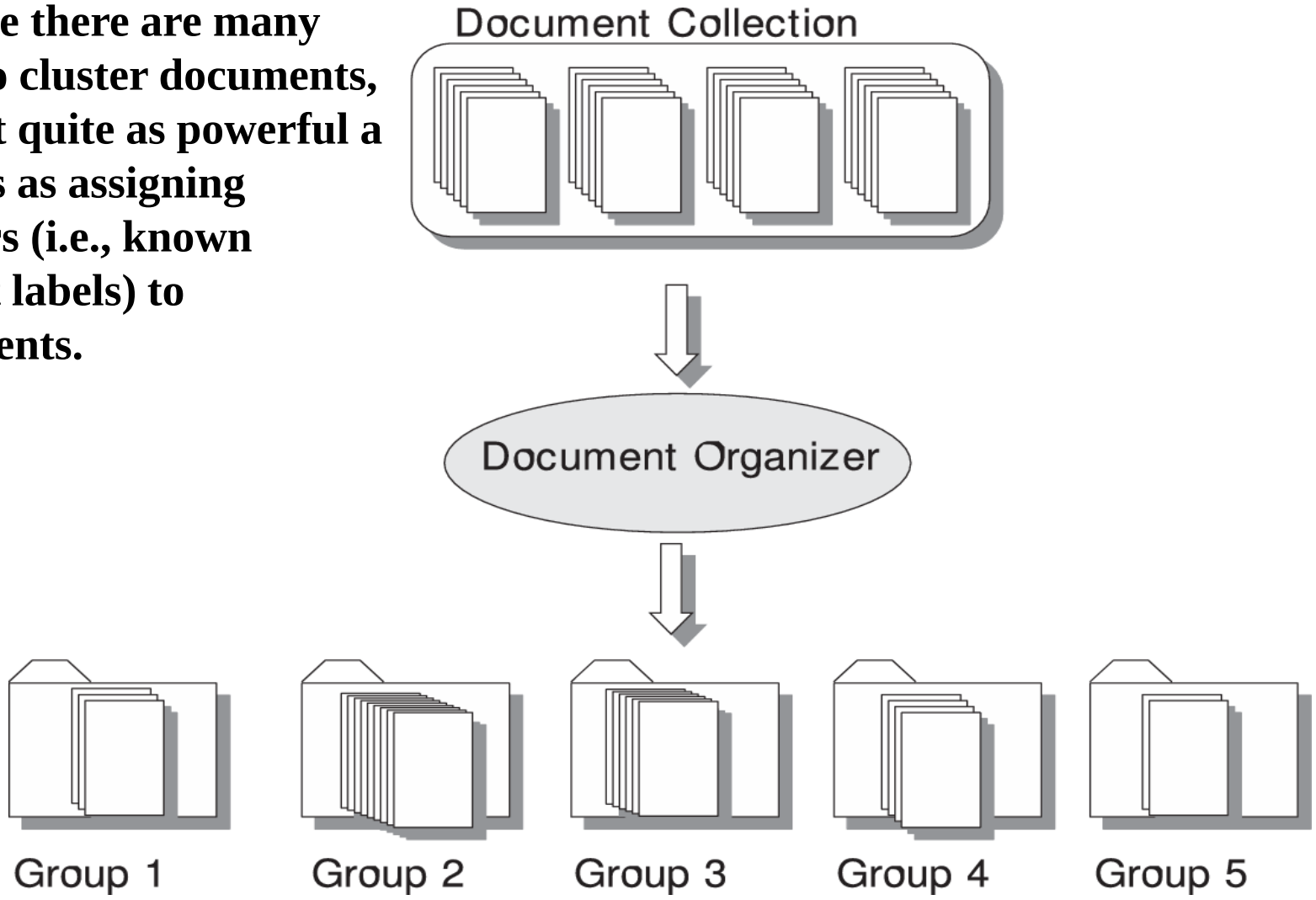
Text Clustering

The clustering process is equivalent to assigning the labels needed for text categorization.



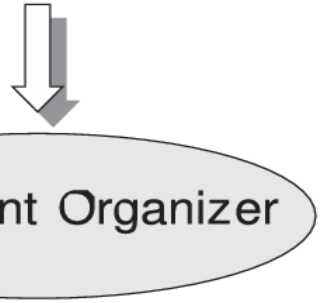
Text Clustering

Because there are many ways to cluster documents, it is not quite as powerful a process as assigning answers (i.e., known correct labels) to documents.



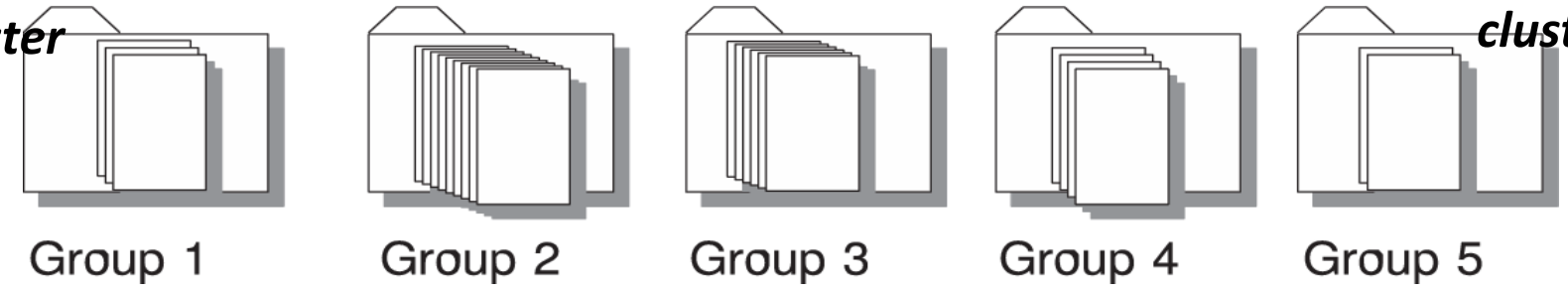
Text Clustering

By studying key words that characterize a cluster, we can gain insights into that cluster.



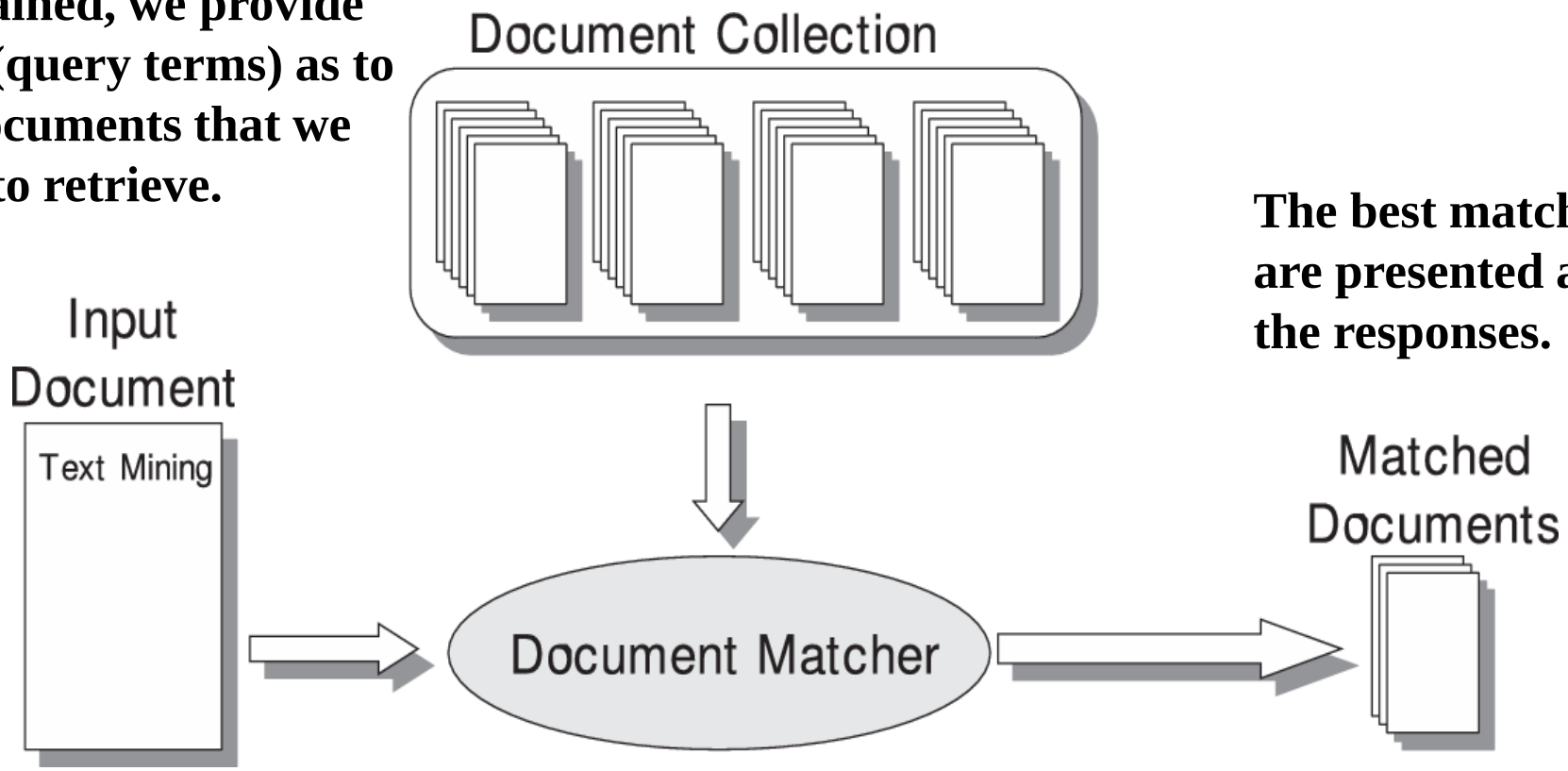
A animal cluster

A health cluster



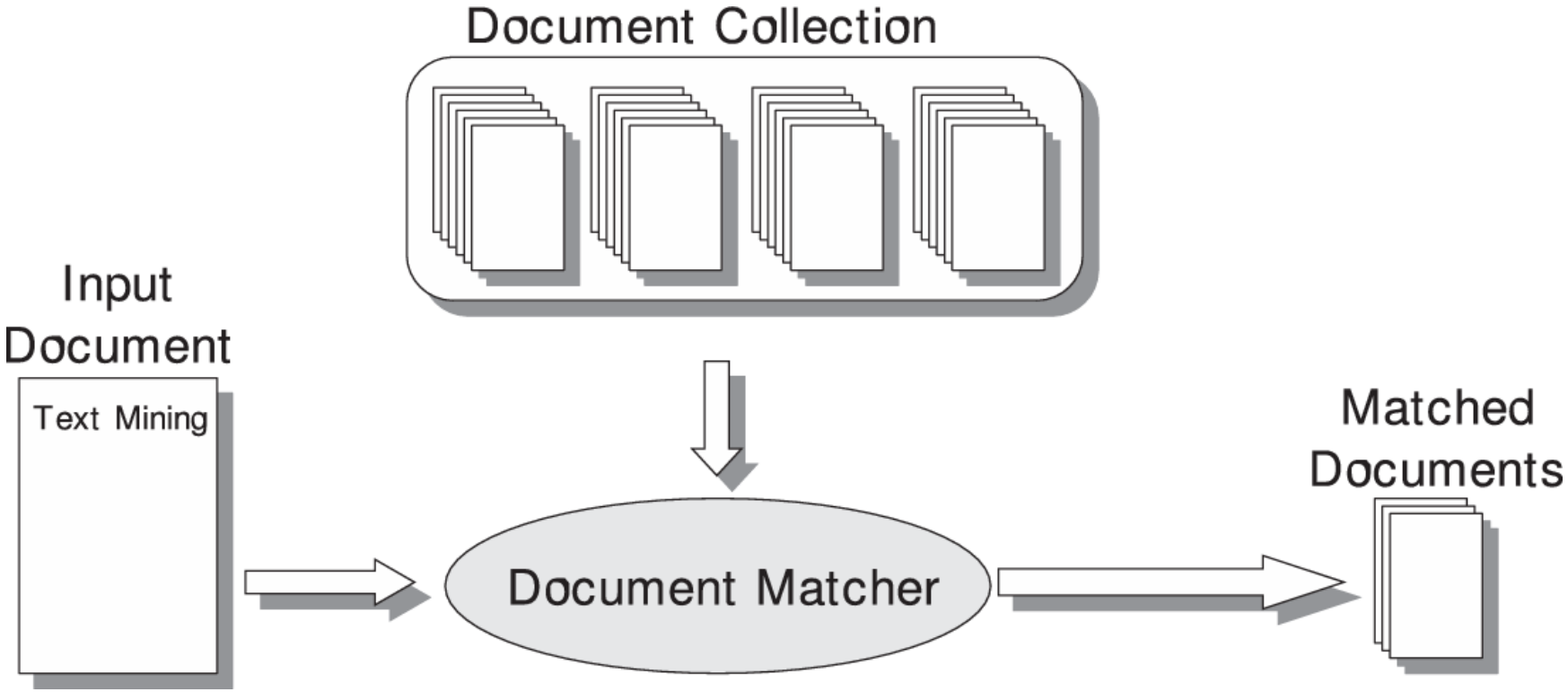
Information Retrieval

A collection of documents is obtained, we provide clues (query terms) as to the documents that we want to retrieve.



The best matches are presented as the responses.

Information Retrieval



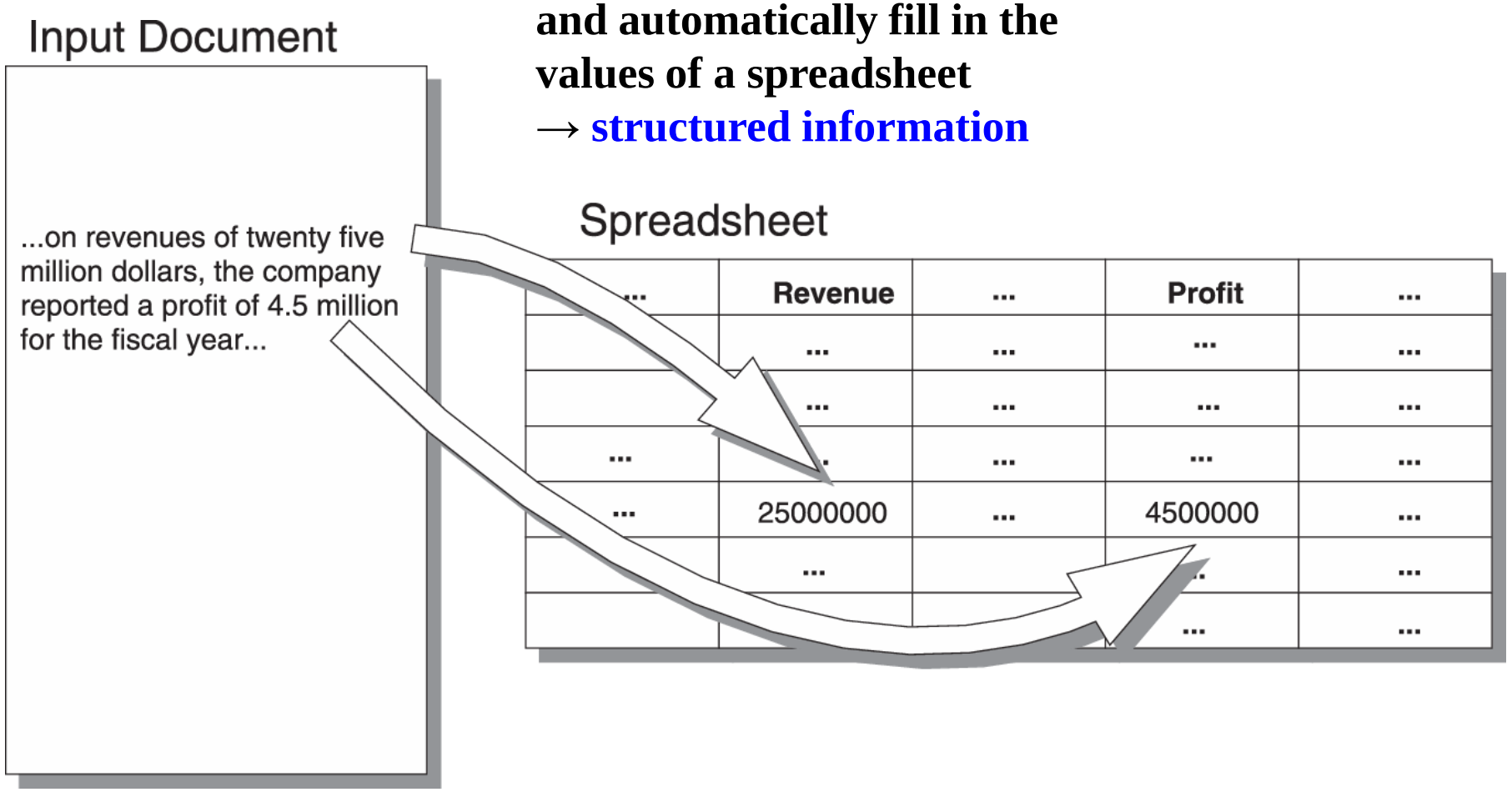
- **The process can be generalized to a document matcher, where instead of a few words, a complete document is presented as a set of clues.**
- **The input document is then matched to all stored documents, retrieving the best-matched documents.**

Information Retrieval

- A basic concept for information retrieval is measuring similarity
  - A comparison is made between two documents, measuring how similar the documents are.
  - For comparison, even a small set of words input into a search engine can be considered as a document that can be matched to others.
  - From one perspective, measuring similarity is related to predictive methods for learning and classification that are called **nearest-neighbor** methods.

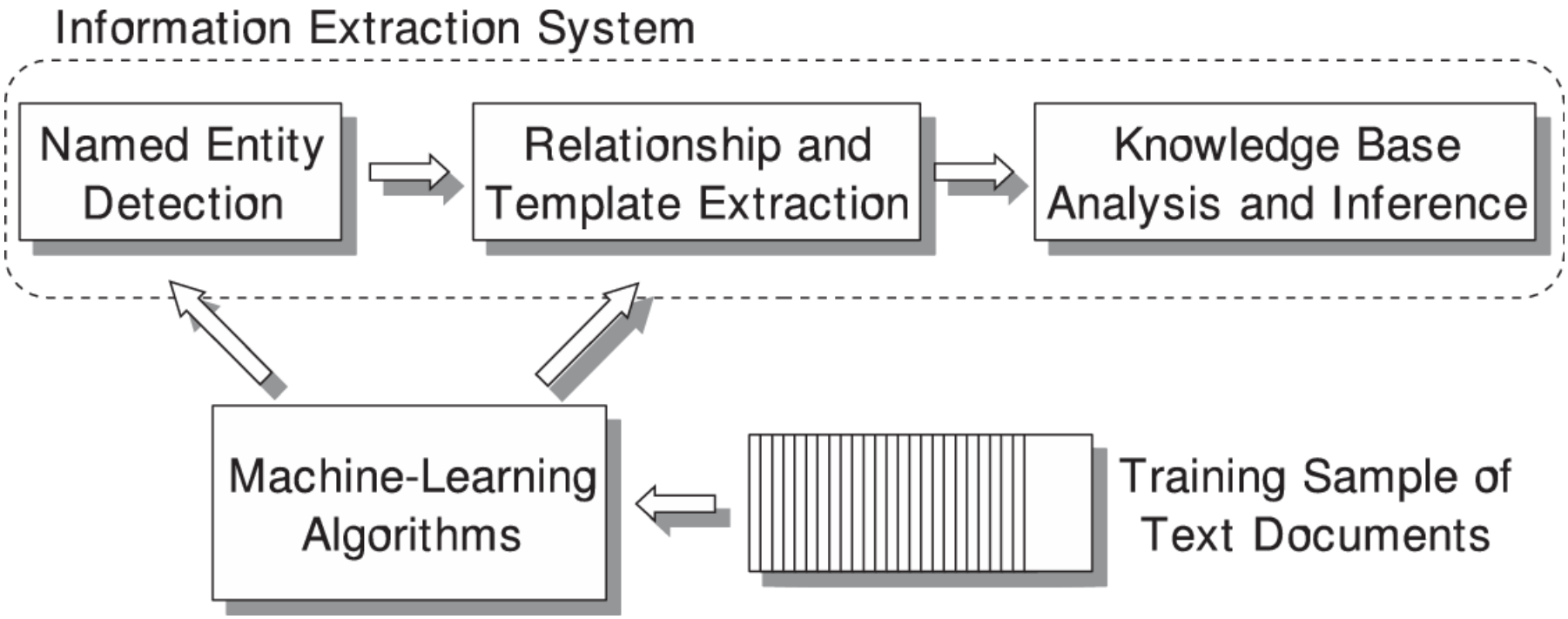
Information Extraction

The objective here is to take an unstructured document and automatically fill in the values of a spreadsheet → structured information



Information Extraction

- Although the most accurate information extraction systems often involve handcrafted language processing modules, substantial progress has been made in applying machine-learning techniques.



One of the many differences between *Robert L. James, chairman and chief executive officer of McCann-Erickson*, and *John J. Dooner, Jr.*, the agency's president and chief operating officer, is quite telling: Mr. James enjoys sailboating, while Mr. Dooner owns a powerboat.

Now, Mr. James is preparing to sail into the sunset, and Mr. Dooner is poised to rev up the engines to guide *Interpublic Group's McCann-Erickson* into the 21st century. Yesterday, *McCann* made official what had been widely anticipated: *Mr. James*, *57* years old, is stepping down as chief executive officer on *July 1* and will retire as chairman at the *end of the year*. He will be succeeded by *Mr. Dooner*, *45* ...

Fig. 6.1 WSJ text with entity mentions emphasized by italic fonts

Table 6.1 Extracted position change information

Organization	*McCann-Erickson*
Position	*Chief executive officer*
Date	*July 1*
Outgoing person name	*Robert L. James*
Outgoing person age	*57*
Incoming person name	*John J. Dooner, Jr.*
Incoming person age	*45*

Automatic Summarization

https://www.google.com/search?hl=en&gl=us&tbm=&authuser

text mining - Google Search

Google

text mining

Edmund

All News Books Images Videos More Search tools

About 60,100,000 results (0.37 seconds)

Text mining, also referred to as **text** data **mining**, roughly equivalent to **text** analytics, refers to the process of deriving high-quality information from **text**. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

Text mining - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Text_mining Wikipedia

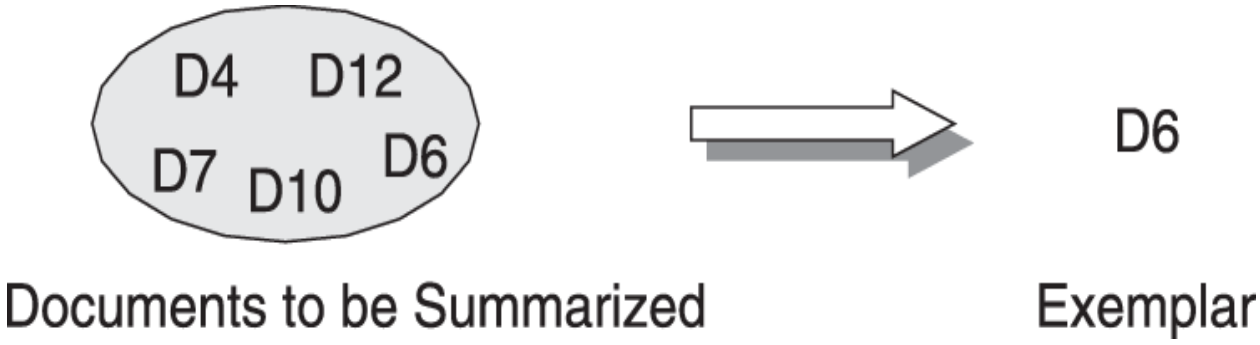
More about Text mining

Feedback

Text mining - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Text_mining Wikipedia
Text mining, also referred to as **text** data **mining**, roughly equivalent to **text** analytics, refers to the process of deriving high-quality information from **text**. High-quality

Automatic Summarization

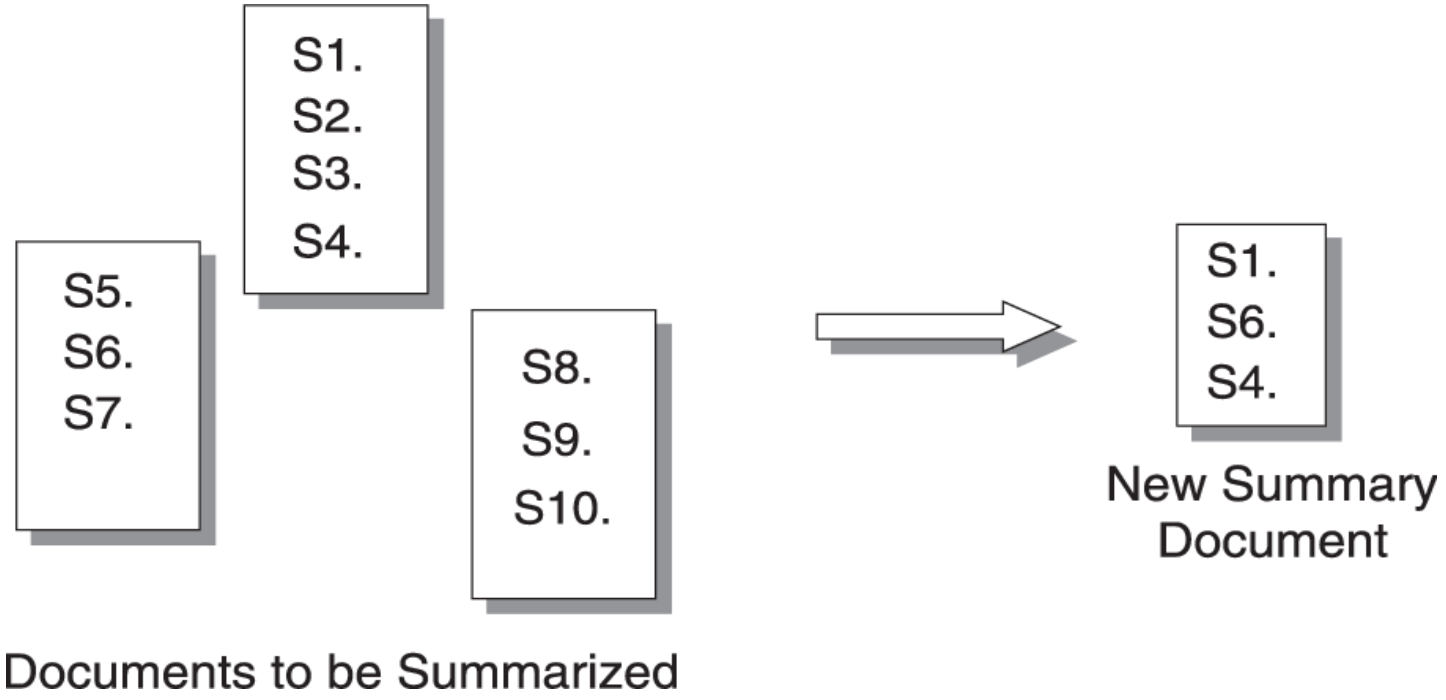
- Automatic summarization from multiple social media sources is a highly active research topic that aims at reducing and aggregating the amount of information presented to users.



- A principal technical approach to summarization is closely aligned with **clustering**.
- A cluster consists of similar documents, about the same topic.
- Pick the most representative document as a summary of the whole cluster of documents.

Automatic Summarization

- Alternatively, we can provide a summary by selecting sentences from the given documents, instead of selecting a single representative document, and merging them into a single summary.



Event Detection

- Event detection in social media texts is important because people tend to post many messages about current events, and many users read those comments in order to find the information that they need.

- Event detection techniques can be classified according to the **event type** (specified or unspecified), the **detection task** (retrospective or new event detection), and the **detection method** (supervised or unsupervised).

TABLE 1. Taxonomy of Event Detection Techniques in Twitter.

References	Type of event		Detection method		Detection task		Application
	Specified	Unspecified	Supervised	Unsupervised	NED	RED	
Sankaranarayanan et al. (2009)		x	x	x	x		Breaking-news detection
Phuvipadawat and Murata (2010)		x		x	x		Breaking-news detection
Petrović et al. (2010)		x		x	x		General (unknown) event detection
Becker et al. (2011a)		x	x	x	x		General (unknown) event detection
Long et al. (2011)		x		x	x		General (unknown) event detection
Weng and Lee (2011)		x		x	x		General (unknown) event detection
Cordeiro (2012)		x		x	x		General (unknown) event detection
Popescu and Pennacchiotti (2010)	x		x		x		Controversial news events about celebrities
Popescu et al. (2011)	x		x		x		Controversial news events about celebrities
Benson et al. (2011)	x		x			x	Musical event detection
Lee and Sumiya (2010)	x			x	x		Geosocial event monitoring
Sakaki et al. (2010)	x		x		x		Natural disaster events monitoring
Becker et al. (2011)	x		x			x	Query-based event retrieval
Massoudi et al. (2011)	x			x		x	Query-based event retrieval
Metzler et al. (2012)	x			x		x	Query-based structured event retrieval
Gu et al. (2011)	x			x		x	Query-based structured event retrieval

Sentiment Analysis: An Overview

Sentiment Analysis

- **Sentiment analysis**, also called **opinion mining**, is the field of study that analyzes people's:

  opinions, sentiments, evaluations, appraisals, attitudes, and emotions

  towards **entities** such as:

  products, services, organizations, individuals, issues, events, topics, and their attributes.

Sentiment Analysis

- The following different names all refer to sentiment analysis, emphasizing on slightly different but related tasks:

  sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc.

- They now all fall under the umbrella term **sentiment analysis**, although some people in the academia still use opinion mining.

Affect

- **Definition (Merriam-Webster)** a set of observable manifestations of a subjectively experienced emotion

  … patients … showed perfectly normal reactions and affects …

- *Effect* and *affect* are often confused because of their similar spelling and pronunciation. The uncommon noun *affect*, which has a meaning relating to psychology, is also sometimes mistakenly used for the very common *effect*. In ordinary use, the noun you will want is *effect* (*waiting for the new law to take effect; the weather had an effect on everyone's mood*).

Sentiment Analysis & NLP

- Although linguistics and natural language processing (NLP) have a long history, little research had been done about people's opinions and sentiments before the year **2000**.
- But since then, sentiment analysis has become a very active research area. There are several reasons for that:
  - It has a wide arrange of applications, almost in every domain. This provides a strong motivation for research.
  - It offers many challenging research problems, which had never been studied before.
  - For the first time in human history, we now have a huge volume of opinionated data in social media.
    - In fact, sentiment analysis is now right at the center of the social media research.

Sentiment Analysis & Social Media

- Social media offers a **World-of-Mouth**
  - **User-generated content** is teeming with people's feelings & emotions
    - Personal experiences and opinions about anything in forums, blogs, Twitter, etc.
    - Postings at social networking sites, e.g., facebook.
    - Comments about articles, issues, topics, reviews, etc.
- Social media offers an ocean of opportunities
  - Social media is on the global scale
  - The ability to survey a large pool of people easily has enormous potential.

An Ocean of Opportunities

- **Businesses** spend big bucks to find consumer opinions using consultants, surveys and focus groups, etc., to benchmark products and services
- **Advertisers** want to know what type of people are willing to purchase their products, and why.
  - **Ads placements**:
    - Place ads in the social media content
    - Place an ad if one praises a product.
    - Place an ad from a competitor if one criticizes a product
  - **Brand Monitoring**

An Ocean of Opportunities

- For individuals
  - **Consumers** want to make more informed decisions to buy products or to use services
  - **Politicians** want to know public opinions (polls) about their policies and their reputation. Traditional polling methods can take days and weeks to deliver results.
  - **Voters** want to know public opinions about political candidates and issues.
  - **Researchers** want to develop systems that could extract valuable data accurately (and publish more papers)

Representative Applications

Products & Sales

- **Mishne & Glance (2006):** showed that positive sentiment is a better predictor of movie success than buzz counts.
- **Liu et al. (2007)**: used sentiment analysis to predict **movie sales**
- **Asur & Huberman (2010), Joshi et al. (2010):** Used Twitter data, movie reviews, and blogs to predict box-office revenues for **movies**.

Predicting Movie Sales from Blogger Sentiment

Gilad Mishne
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
gilad@science.uva.nl

Natalie Glance
Intelliseek Applied Research Center
5001 Baum Blvd, Pittsburgh, PA 15213
nglance@intelliseek.com

They showed that positive sentiment is a better predictor of movie success than buzz counts.

Abstract

The volume of discussion about a product in weblogs has recently been shown to correlate with the product's financial performance. In this paper, we study whether applying sentiment analysis methods to weblog data results in better correlation than volume only, in the domain of movies. Our main finding is that positive sentiment is indeed a better predictor for movie success when applied to a limited context around references to the movie in weblogs, posted prior to its release.

If my film makes one more person miserable, I've done my job.
– Woody Allen

Introduction

Weblogs provide online forums for discussion that record the voice of the public. Woven into this mass of discussion his unhappiness with Dell as a company reverberated across the blogosphere and into the press, creating a public relations mini-fiasco for Dell.

Hard evidence of the influence of online discussion on consumer decisions is beginning to emerge. An Intelliseek survey of 660 online consumers showed that people are 50 percent more likely to be influenced by word-of-mouth recommendations from their peers than by radio/TV ads [3]. Researchers at IBM reported that online blog postings can successfully predict spikes in the sales rank of books (Gruhl, Guha, Kumar, Novak, & Tomkins, 2005), showing that the raw number of posts about a book was a good predictor.

However, opinion comes in many flavors: positive, negative, mixed, and neutral mixed in with splashes of sarcasm, wit and irony. Novel techniques in sentiment analysis make it possible to quantify the aggregate level of positive vs. neg-

apparently an early easter is bad for apparel sales. who knew? i'll probably go see "guess who?" this weekend. i liked miss congeniality but the sequel [link to IMDB's page for "Miss Congeniality 2"] looks *awful*. and seattle's too much of a backwater to be showing D.E.B.S. i had to wait forever to see saved! too. mikalah gordon got kicked off american idol last night. while she wasn't the best singer, i wish …
Monday, March 28, 2005 - Miss Congeniality 2: Armed and Fabulous. I know this is overdue, but I wanted to use this opportunity to discuss an important topic. The issue at hand is known as the Sandra Bullock Effect (SBE). This theorem was first proposed by my brother, Arthur, so he is the real expert, but I will attempt to explain it here. The SBE is the degree to which any movie becomes watchable simply by the presence of a particular actor or actress who you happen to be fond of. For example, if I told you that someone made a movie about a clumsy, socially awkward, dowdy female police officer who goes undercover as a beauty pageant contestant to foil some impenetrable criminal conspiracy, you'd probably think to yourself, "Wow that sounds pretty dumb." And you'd be right. However…

Table 4: Typical references to movies in blogs: pre-release (top), and post-release (bottom).

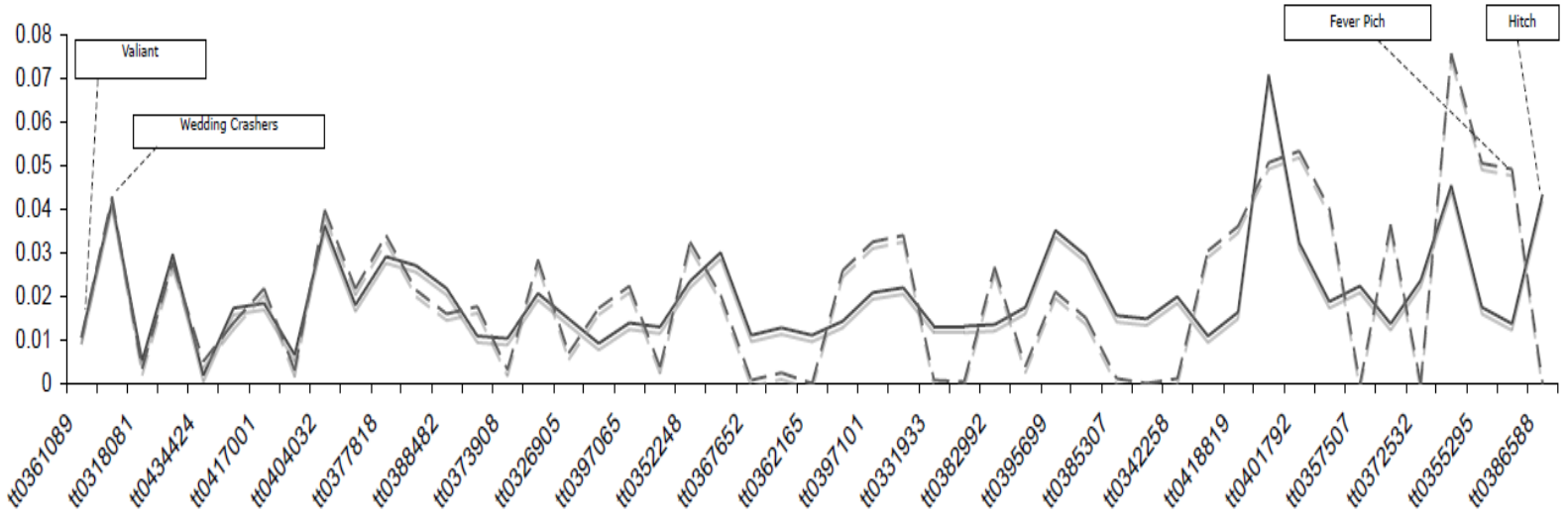


Figure 2: Per-movie comparison of income per screen (blue, continuous line) and positive references (green, dashed line), sorted by degree of correlation. For space reasons, the X-axis shows only the movie IMDB ID.

ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs

Yang Liu[1], Xiangji Huang[2], Aijun An[1] and Xiaohui Yu[2]
[1]Department of Computer Science and Engineering
York University, Toronto, Canada
[2]School of Information Technology
York University, Toronto, Canada
yliu@cse.yorku.ca, jhuang@yorku.ca, aan@cse.yorku.ca, xhyu@yorku.ca

ABSTRACT

Due to its high popularity, Weblogs (or blogs in short) present a wealth of information that can be very helpful in assessing the general public's sentiments and opinions. In this paper, we study the problem of mining sentiment information from blogs and investigate ways to use such information for predicting product sales performance. Based on an analysis of the complex nature of sentiments, we propose Sentiment PLSA (S-PLSA), in which a blog entry is viewed as a

commentaries or discussions on a particular subject, ranging from mainstream topics (e.g., food, music, products, politics, etc.), to highly personal interests [13]. Since many bloggers choose to express their opinions online, blogs serve as an excellent indicator of public sentiments and opinions.

This paper studies the predictive power of opinions and sentiments expressed in blogs. We focus on the blogs that contain reviews on products. Since what the general public thinks of a product can no doubt influence how good it sells, understanding the opinions and sentiments expressed

To better illustrate the effects of the parameter values on the prediction accuracy, we present in Figure 3 the experimental results on a particular movie, *Little Man*. For each parameter, we plot the predicted box office revenues and the true values for each day using different values of the parameter. It is evident from the plots that the responses to each parameter are similar to what is observed from Figure 2. Also note that the predicted values using the optimal parameter settings are close to the true values accross the whole time span. Similar results are also observed on other movies, demonstrating the consistency of the proposed approach for different days.

6.3 Comparison with alternative methods

To verify that the sentiment information captured by the S-PLSA model plays an important role in box office revenue prediction, we compare ARSA with two alternative methods which do not take sentiment information into consideration.

We first conduct experiments to compare ARSA against the pure autoregressive (AR) model without any terms on sentiments, i.e., $y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$. The results are shown in Figure 4. We observe the behaviors of the two models as p ranges from 3 to 7. Apparently, although the accuracy of both methods improves with increasing p, ARSA constantly outperforms the AR model by a factor of 2 to 3.

We then proceed to compare ARSA with an autoregressive model that factors in the volume of blog mentions in prediction. In Section 3, we have illustrated the characteris-

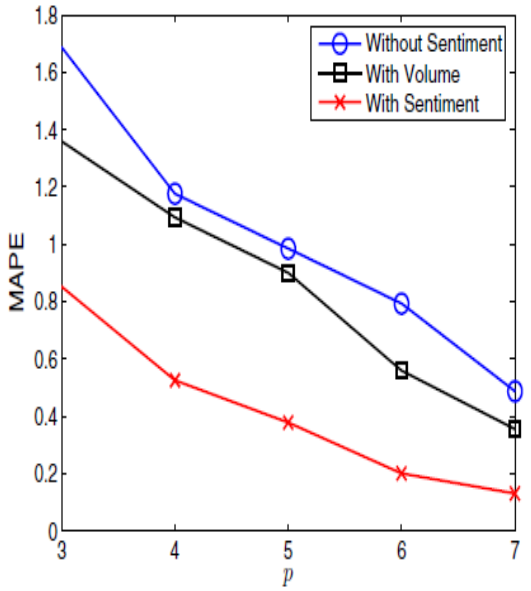


Figure 4: ARSA vs. alternative methods

this, we experiment with the following autoregressive model that utilizes the volume of blogs mentions. In contrast to ARSA, where we use a multi-dimensional probability vector produced by S-PLSA to represent bloggers' sentiments, this model uses a scalar (number of blog mentions) to indicate the degree of popularity. The model can be formulated as

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \rho_i v_{t-i} + \epsilon_t,$$

where y_t's are obtained in the same way as in ARSA, v_{t-i} denotes the number of blog mentions on day $t-i$, and ϕ_i and ρ_i are parameters to be learned. This model can be trained using a procedure similar to what is used for ARSA. Using

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

Used Twitter data, movie reviews, and blogs to predict box-office revenues for movies

Abstract—**In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.**

I. Introduction

Social media has exploded as a category of online discourse

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

We have focused on movies in this study for two main reasons.

- The topic of movies is of considerable interest among the social media user community, characterized both by large number of users discussing movies, as well as a

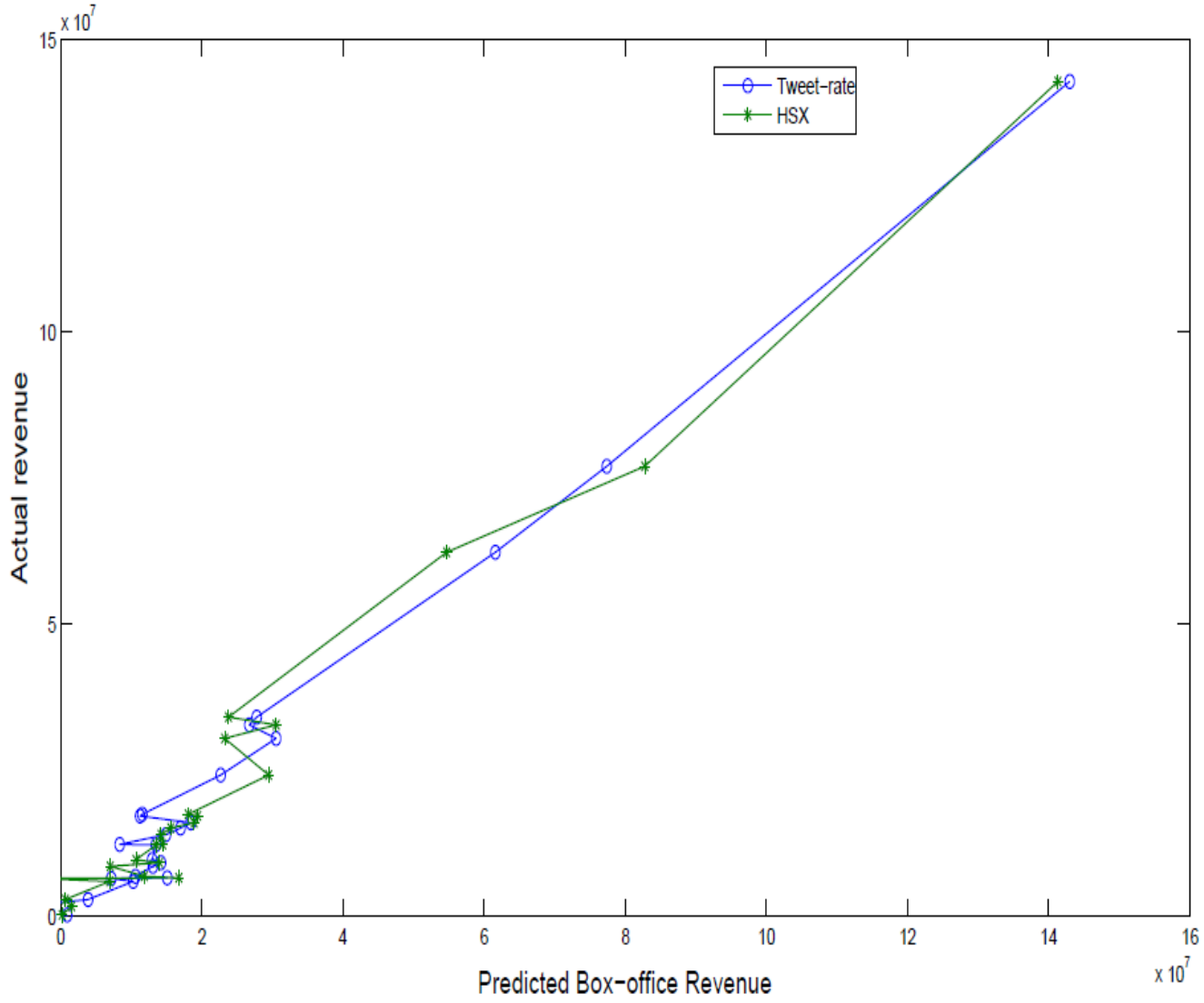


Fig. 6. Predicted vs Actual box office scores using tweet-rate and HSX predictors

Hollywood Stock Exchange

From Wikipedia, the free encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.

Find sources: "Hollywood Stock Exchange" – news · newspapers · books · scholar · JSTOR *(July 2008) (Learn how and when to remove this template message)*

The **Hollywood Stock Exchange**, or **HSX**, is a web-based, multiplayer game in which players use simulated money to buy and sell "shares" of actors, directors, upcoming films, and film-related options.[1] The game uses Virtual Specialist technology invented by HSX co-founders and creators Max Keiser and Michael R. Burns, who were awarded a U.S. patent no. 5950176 in 1999 for the invention. Claims of this patent cover trading applications for trading virtual securities using virtual currencies over a network.

The company moved into the former Ritts Furniture building designed by Harry Harrison on Santa Monica Boulevard.

Contents [hide]

Operation [edit]

Because trading directly affects the prices of the securities – purchasing enough shares of a stock causes its price to rise, and selling causes its price to fall – and because the ultimate value of a moviestock is based on the film's box office, stock prices act as box office predictions. For example, if a particular moviestock trades at

Movie Reviews and Revenues: An Experiment in Text Regression*

Mahesh Joshi Dipanjan Das Kevin Gimpel Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{maheshj,dipanjan,kgimpel,nasmith}@cs.cmu.edu

Abstract

We consider the problem of predicting a movie's opening weekend revenue. Previous work on this problem has used metadata about a movie—e.g., its genre, MPAA rating, and cast—with very limited work making use of text *about* the movie. In this paper, we use the text of film critics' reviews from several sources to predict opening weekend revenue. We describe a new dataset pairing movie reviews with metadata and revenue data, and show that review text can substitute for metadata, and even improve over it, for prediction.

1 Introduction

correlation between actual revenue and sentiment-based metrics, as compared to mention counts of the movie. (They did not frame the task as a revenue prediction problem.) Zhang and Skiena (2009) used a news aggregation system to identify entities and obtain domain-specific sentiment for each entity in several domains. They used the aggregate sentiment scores and mention counts of each movie in news articles as predictors.

While there has been substantial prior work on using critics' reviews, to our knowledge all of this work has used polarity of the review or the number of stars given to it by a critic, rather than the review text directly (Terry et al., 2005).

Our task is related to sentiment analysis (Pang et

the dependency relation features (set III) to the n-grams does improve the performance enough to make it significantly better than the metadata-only baseline for per screen revenue prediction.

Salient Text Features: Table 3 lists some of the highly weighted features, which we have categorized manually. The features are from the text-only model annotated in Table 2 (total, not per screen). The feature weights can be directly interpreted as U.S. dollars contributed to the predicted value $\hat{y}$ by each occurrence of the feature. Sentiment-related features are not as prominent as might be expected, and their overall proportion in the set of features with non-zero weights is quite small (estimated in preliminary trials at less than 15%). Phrases that refer to metadata are the more highly weighted and frequent ones. Consistent with previous research, we found some positively-oriented sentiment features to be predictive. Some other prominent features not listed in the table correspond to special effects ("*Boston Globe*: of_the_art", "and_cgi"), particular movie franchises ("shrek_movies", "*Variety*: chronicle_of", "voldemort"), hype/expectations ("blockbuster", "anticipation"), film festival ("*Variety*: canne" with negative weight) and time of re-

	Feature	Weight ($M)
rating	pg	+0.085
	New York Times: adult	-0.236
	New York Times: rate_r	-0.364
sequels	this_series	+13.925
	LA Times: the_franchise	+5.112
	Variety: the_sequel	+4.224
people	*Boston Globe*: will_smith	+2.560
	Variety: brittany	+1.128
	^_producer_brian	+0.486
genre	*Variety*: testosterone	+1.945
	Ent. Weekly: comedy_for	+1.143
	Variety: a_horror	+0.595
	documentary	-0.037
	independent	-0.127
sentiment	*Boston Globe*: best_parts_of	+1.462
	Boston Globe: smart_enough	+1.449
	LA Times: a_good_thing	+1.117
	shame_$	-0.098
	bogeyman	-0.689
plot	*Variety*: torso	+9.054
	vehicle_in	+5.827
	superhero_$	+2.020

Table 3: Highly weighted features categorized manually. ^ and $ denote sentence boundaries. "brittany" frequently refers to Brittany Snow and Brittany Murphy. "^_producer_brian" refers to producer Brian Grazer (*The Da Vinci Code*, among others).

Proceedings of the Third International ICWSM Conference (2009)

They used sentiment analysis to measure **social influences in online book reviews:** How existing reviews affect new reviews?

Analysis of Social Influence in Online Book Reviews

Patty Sakunkoo and Nathan Sakunkoo

Stanford University, Stanford, CA 94305
{psak, sakunkoo} @ stanford.edu

Abstract

It has been widely recognized that online opinions constitute important informational sources for consumers and producers. The open nature of communication supported by social media, however, raises an important yet unsettled question of whether and how earlier opinions affect those that come after. This poster presents a model to illustrate the relationship between existing and new reviews. Based on 12,500 Amazon reviews, our choice model shows support for the idea that social contagion may be an important mechanism guiding behaviors of online reviewers. The results thus offer novel insights toward a better understanding of contagious behaviors, as well as minority influence, among social media users.

As part of the larger effort to examine the interactions among social media users and how opinions evolve, this poster explores a new interpretation by developing and testing a model for socially-influenced online reviews. The results challenge existing theoretical conjectures above by providing an illustration of how the observed trend can be explained by a *contagious* tendency.

Methods

For preliminary investigation of the aggregate patterns, a simulation was conducted to test whether the existing opinions may prompt or provide bases for a potential reviewer to write a new review. A result from this

Representative Applications

Politics

- O'Connor et al. (2010), **From Tweets to Polls**: tracked how Twitter sentiment was linked with public opinion polls
- Tumasjan et al. (2010): tracked Twitter sentiment to predict election results
- Chen et al. (2010): used sentiment analysis to study political standpoints

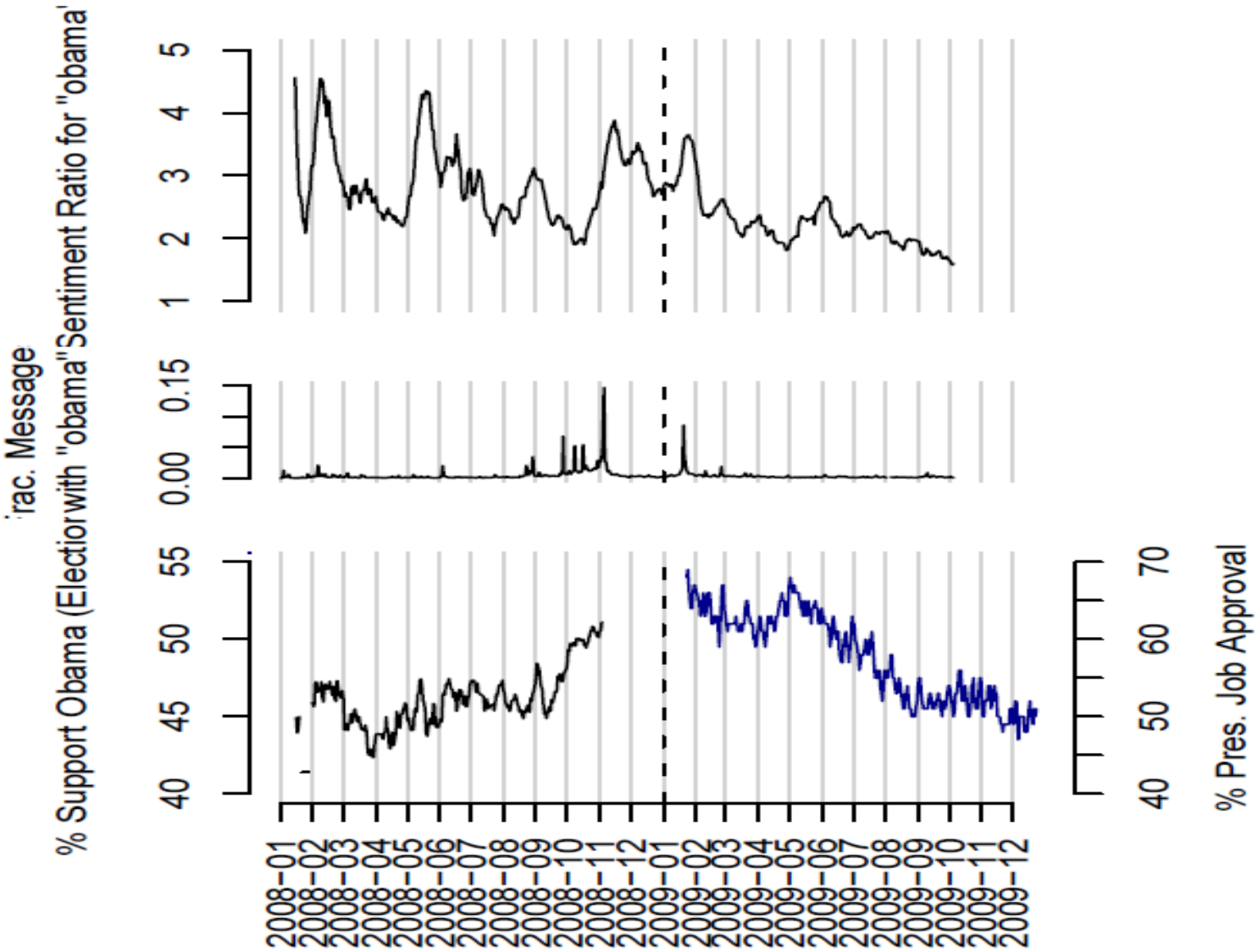


Figure 9: The sentiment ratio for *obama* (15-day window), and fraction of all Twitter messages containing *obama* (day-by-day, no smoothing), compared to election polls (2008) and job approval polls (2009).

Representative Applications

Stock Market

- Bollen et al. (2011), **Twitter mood predicts the stock market** : used Twitter moods to predict the stock market.
  - Public mood states along 7 different dimensions of mood are measured from the text content of large-scale Twitter feeds (342,255) tweets. (Next slide)
  - Daily variations in public mood states show statistically significant correlation to daily changes in Dow Jones Industrial Average closing values.
  - Certain dimensions of public mood states, in particular Calm, increase the accuracy of a Self Organizing Fuzzy Neural Network model in predicting up and down changes in DJIA closing values to 87.6%.
- Bar-Haim et al. (2011): used sentiment analysis to identify **expert investors** (vs. non-expert)

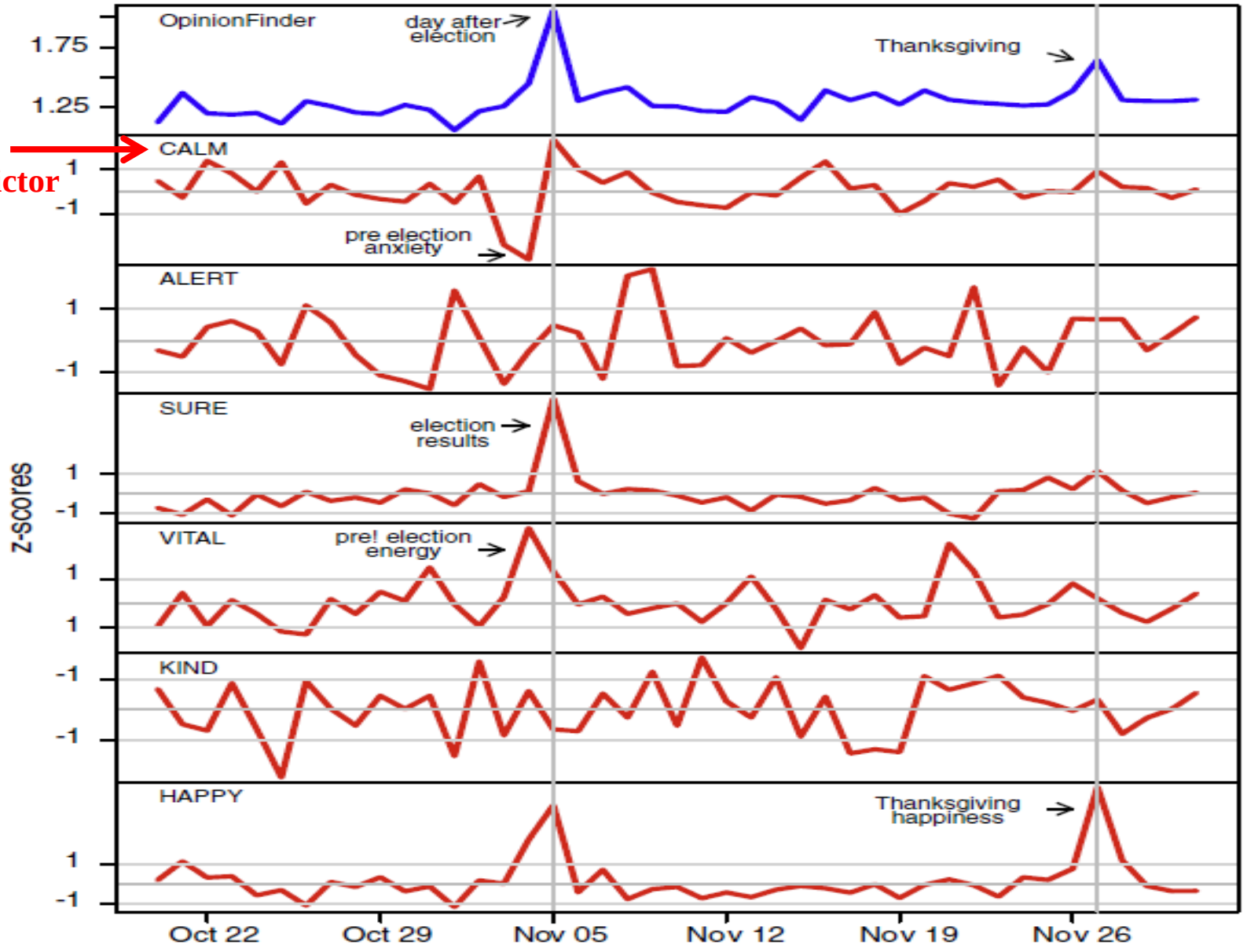


Fig. 2. Tracking public mood states from tweets posted between October 2008 to December 2008 shows public responses to presidential election and thanksgiving.

Google-Profile of Mood States (GPOMS) – 6 dimensions

Representative Applications

Stock Market

- Zhang and Skiena (2010), **Trading Strategies To Exploit Blog and News Sentiment**: used blog and news sentiment to study **trading strategies**
  - Their findings provided them with a sentiment-based trading strategy which gives consistently favorable returns with low volatility over a five year period (2005-2009).

StockTwits

AAPL Apple Inc. — Stock Price a

https://stocktwits.com/symbol/AAPL

Stocktwits Rooms Rankings Earnings Calendar Shop

Symbol or @Username

Post

DOW 0.18% S&P 500 0.30% NASDAQ 0.47% Trending now BTC.X 6.65% SNAP 32.30% JKS 9.39% NFLX 6.73% PYPL 4.58% T 0.37% PINS 10.78% FB 4.81% ATNM

Watchlist My Rooms Today

Sort by My sort Done

Search to add

Add symbols to your watchlist for easy access to your favorite stocks

Apple Inc.

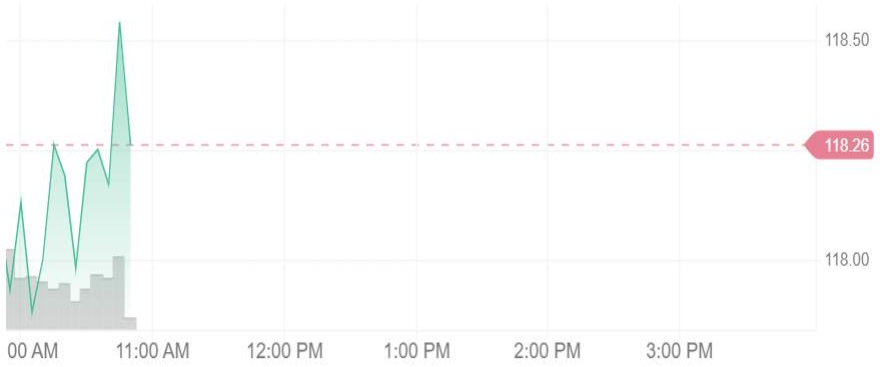
519,746

Watch

AAPL 118.31 ↑ 0.80 (0.68%)

NASDAQ Updated Oct 21, 2020 10:50 AM

↑ 0.68%	↑ 0.22%	↓ 36.50%
Price	Sentiment	Message Volume



1d 1w 1m 3m 6m 1y All

FEATURED VIDEOS Powered by [primis]

Biden's Lead Is Not Insurmountable,...

Oct.19 -- Greg Valliere, chief U.S. policy strategist at AGF Investments, says Joe Biden's lead in the polls is not insurmountable and he doesn't think a stimulus deal will...

Apple Inc. 118.23 ↑ 0.72 (0.61%) Watch

Options500to100k Bullish now
$AAPL new phones new games .. more games .. more kids will use momma credit card to buy games in there. A growing technology.. it also expanding all over the world.. India .. even Syria! All times highs will come $QQQ

jcdoza Bullish now
$AAPL good swing trade to 150 by EOY.

STOption now
$AAPL great support @118.2 boy!

umsyed Bullish now
$AAPL show me that $119 quickly

News

More News

A new Analyst Report sees Apple's Push into 5G a Key Driver for Growth while…
09:12 AM - Patently Apple

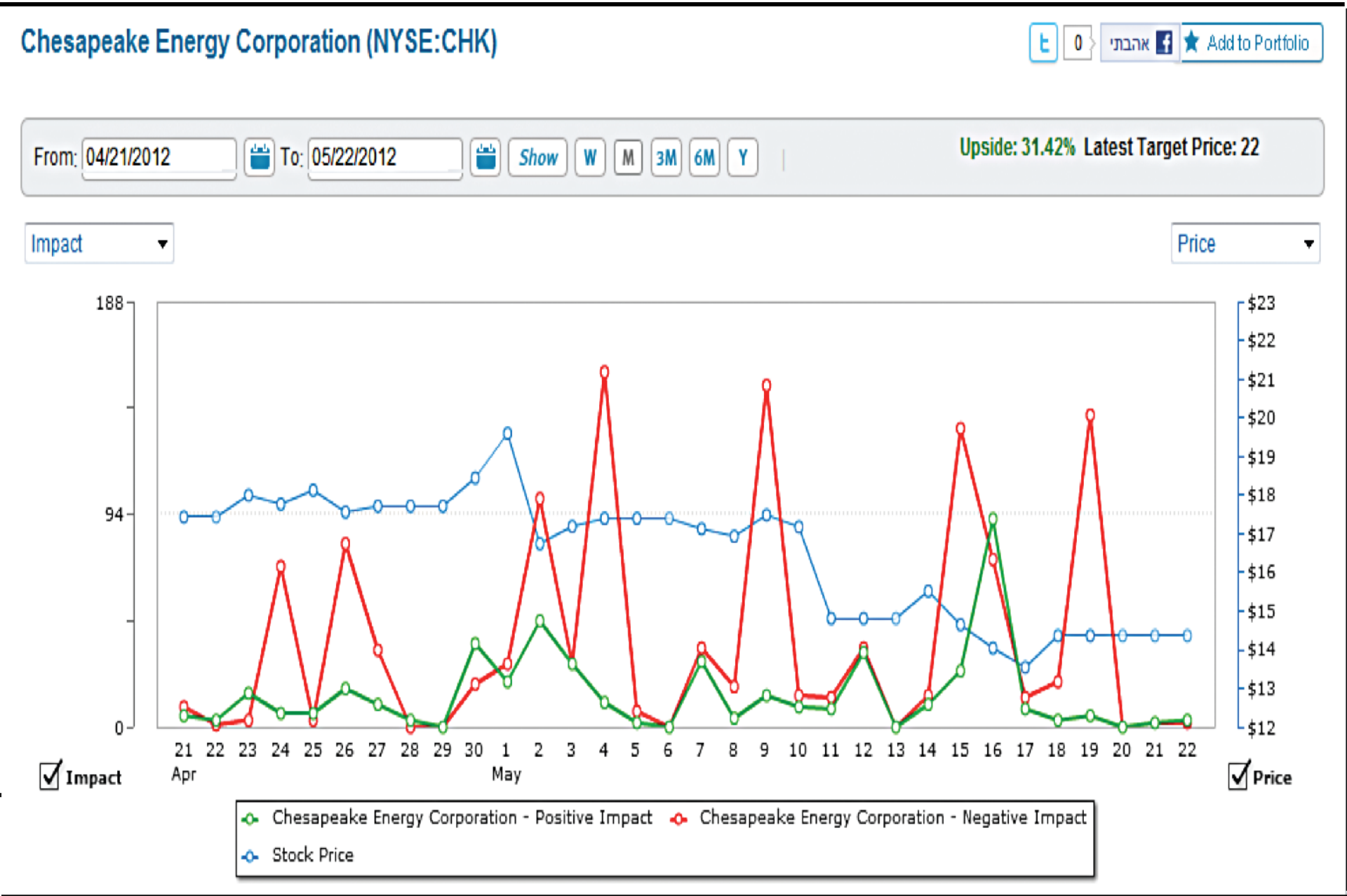
iPad Air (2020) review: A tablet designed for work and play
09:00 AM - ZDNet Latest News

Apple iPad Air (2020) review: take it from the Pro
09:00 AM - The Verge

Get The Stocktwits Daily Rip

Enter your email Subscribe

Sentiment Graph (http://www.thestocksonar.com).



The Stock Sonar - Sentiment Service of US Stocks

Data

Published by: Digital Trowel Inc.

Categories: Capital Markets, Business and Finance

Date added: 10/10/2011

Get support for this offering

The Stock Sonar Sentiment Service provides sentiment scores for public companies trades on the US stock market. The Stock Sonar retrieves, reads and analyzes information from a wide variety of online sources including articles, blogs, press releases and other publicly available information based on an in-depth understanding of a text's meaning. The platform's advanced capabilities enable intelligence discerning of positive and negative nuances and events quantifying them to provide investors with immediate insights. The Stock Sonar Sentiment Service presents the following data for each public company: Daily Sentiment Score, Daily Positive Score, Daily Negative Score. The Daily Sentiment Score is a figure between -1 and 1 (where -1 is most negative and 1 is most positive). The Daily Sentiment Score is a weighted average of that day's news articles sentiments. The Daily Positive Score is the weighted sum of all

500 Transactions/month	$0.00 per month
1,000 Transactions/month	$30.00 per month
5,000 Transactions/month	$129.00 per month
25,000 Transactions/month	$599.00 per month
50,000 Transactions/month	$999.00 per month

Service documentation
Service client developer guidelines

RESOURCES

Microsoft PowerPivot For Excel 2010 ▶
Learn how to use Microsoft PowerPivot for Excel 2010, with this and other DataMarket data, to create compelling self-service BI solutions.

The Stock Sonar — Sentiment Analysis of Stocks Based on a Hybrid Approach

[1,2]Ronen Feldman, [2]Benjamin Rosenfeld, [2]Roy Bar-Haim and [2]Moshe Fresko
[1]School of Business Administration, The Hebrew University of Jerusalem, Jerusalem, ISRAEL
[2]Digital Trowel, Airport City, ISRAEL
ronen.feldman@huji.ac.il,{grur,roy,moshe}@digitaltrowel.com

Abstract

The Stock Sonar (TSS) is a stock sentiment analysis application based on a novel hybrid approach. While previous work focused on document level sentiment classification, or extracted only generic sentiment at the phrase level, TSS integrates sentiment dictionaries, phrase-level compositional patterns, and predicate-level semantic events. TSS generates precise in-text sentiment tagging as well as sentiment-oriented event summaries for a given stock, which are also aggregated into sentiment scores. Hence, TSS allows investors to get the essence of thousands of articles every day and may help them to make timely, informed trading decisions. The extracted sentiment is also shown to improve the accuracy of an existing document-level sentiment classifier.

trast, TSS provides precise *sentiment extraction*: highlighting of positive and negative expressions within the article text, as well as extraction of positive and negative business events. Extracted sentiment provides the user an *explanation* for the article score as well as an effective *summary* of multiple news articles. As we show in this paper, the sentiment extracted by TSS can also be used to improve document-level sentiment classifiers.

Another limitation of previous methods is concerned with the level of linguistic analysis required to correctly predict sentiment. Current systems usually employ sentiment lexicons and machine-learning algorithms that operate at the word or phrase level. Such methods typically fail to model compositional expressions, e.g. correctly classifying *"reducing losses"* as positive, but *"reducing forecasts"* as negative. Furthermore, it is often necessary to go beyond the

2 Architecture

TSS collects thousands of articles from thousands of sources every day. The articles are collected via stock-specific RSS feeds, so that each article is associated with one or more stocks. The same news is often repeated by multiple sources. Currently we do not attempt to identify these duplicates, except for the obvious case where articles have the same title and date. However, we assume that the number of times a story is repeated is indicative for its significance, and therefore keeping these duplicates is beneficial.

The collected articles are first cleaned so that the main body of the article is maintained and the extraneous content (such as ads, links to other stories, etc.) is deleted. The module in charge of the extraction of the main textual content from the HTML pages is based on a supervised machine learning approach and a visual training module as described in (Rosenfeld, Feldman, and Ungar 2008). The output of this module is plain text. Each article is analyzed separately for each of its associated stocks. We shall refer to the stock for which the article is currently analyzed as the *main company*. Based on the extracted sentiment, an article score is computed, and article scores are then aggregated into daily scores for each ticker. The rest of this section details sentiment extraction and scoring.

Sentiments		
	Positive	**Negative**
Adjectives	attractive, superior	inefficient, risky
Verbs	invents, advancing	failed, lost
Nouns	opportunity, success	weakness, crisis
Multi-word expressions	exceeding expectations, falling into place	chapter 11,pull back
Neutral expressions	in the worst case, best practice	
Sentiments Modifiers		
Emphasis	Huge, incredible, highly	
De-emphasis	mostly, quite	
Reversal	far from, cut, no	

Table 1: Lexicon components

evaluating the sentiment of each token. Furthermore, because the three components operate within the CARE framework, they can utilize its CRF classifier, which flexibly connects them to the state of the art NER and POS taggers.

The TSS rulebook was developed by a team of three linguistic engineers, assisted by two financially-trained domain experts, over a period of five months. We now turn to a short description of its individual components.

Category	Example (polarity)
Legal	ArvinMeritor (ARM), a maker of integrated systems, rose after it won an antitrust suit against electrical power gear maker **Eaton (ETN)**. ⊖
Analyst Recommendation	On June 23, Caris & Co. reiterated its "buy" rating on **CRM** and increased its price target to $115 from $100. ⊕
Financial	**Western Digital** earnings climb 35% ⊕
Stock Price Change	**SandRidge Energy** fell 3.0 percent to $6.24. ⊖
Deals	The U.S. Army's Mission Installation Contracting Command has awarded **Northrop Grumman Corporation (NYSE:NOC)** a contract to provide logistics support at Fort Eustis, Va. ⊕
Mergers and Acquisitions	On Monday, Ramius LLC has offered to acquire all the outstanding shares of **CYPB** for $4.00 per share in cash. ⊕
Partnerships	Stennis has partnered with **Orbital Sciences Corporation** to test the AJ 26 engines that will power the first stage of the company's Taurus II space launch vehicle. ⊕
Product	**The fast-food giant** recalled 12 mil drinking glasses that contain cadmium ⊖
Employment	**Heidrick & Struggles** Appoints New Hedge Fund Leadership Team. ⊕

Table 2: Types of extracted events. Main company is marked in bold. ⊕/⊖ represent positive/negative polarity.

nounced today its new treatment for multiple sclerosis was found effective in clinical trials". It defines in (1) a positive product event. The primary "anchors" these rules utilize are *"treatment"* nouns (5), followed by a positive adjective (6).

2.5 Sentiment Relevance

When analyzing sentiment for a given company, it is crucial to assess that the sentiment indeed refers to that company. O'Hare et al. (2009) suggested to consider only a window of N words around each mention of the main company in the article, and showed that it improves polarity prediction as compared to considering the whole article for the company. Our experiments on a training corpus confirmed that identification of relevant sections is crucial to obtain reasonable precision, and that the distance from a mention of the main company is a good predictor of relevance. However, we found two additional cues for relevant that were not considered by O'Hare at el.:

1. *Directionality* - sentiments that appear after the main company are more likely to be relevant than sentiments preceding it.
2. *Other entities* - entities that appear between the main company mention and the sentiment are good indicators for irrelevance.

Eventually, we implemented a relevance strategy that considered only sentiments that appear after the main company

Representative Applications

- Hong and Skiena (2010), **The Wisdom of Bookies? Sentiment Analysis vs. the NFL Point Spread**: tracked the relationships between the NFL betting line and public opinions in blogs and Twitter
  - The American Football betting market provides a particularly attractive domain to study the nexus between public sentiment and the wisdom of crowds.
  - Based on the text data from LiveJournal blogs, RSS blog feeds captured by Spinn3r, Twitter, and traditional news media.
  - A strategy based on their finding, betting roughly **30 games per year,** identified winner roughly **60%** of the time from 2006 to 2009, well beyond what is needed to overcome the bookie's typical commission (53%)

Mohammad (2011), **From Once Upon a Time to Happily Ever After**: tracked emotions in Brothers Grimm fairy tales

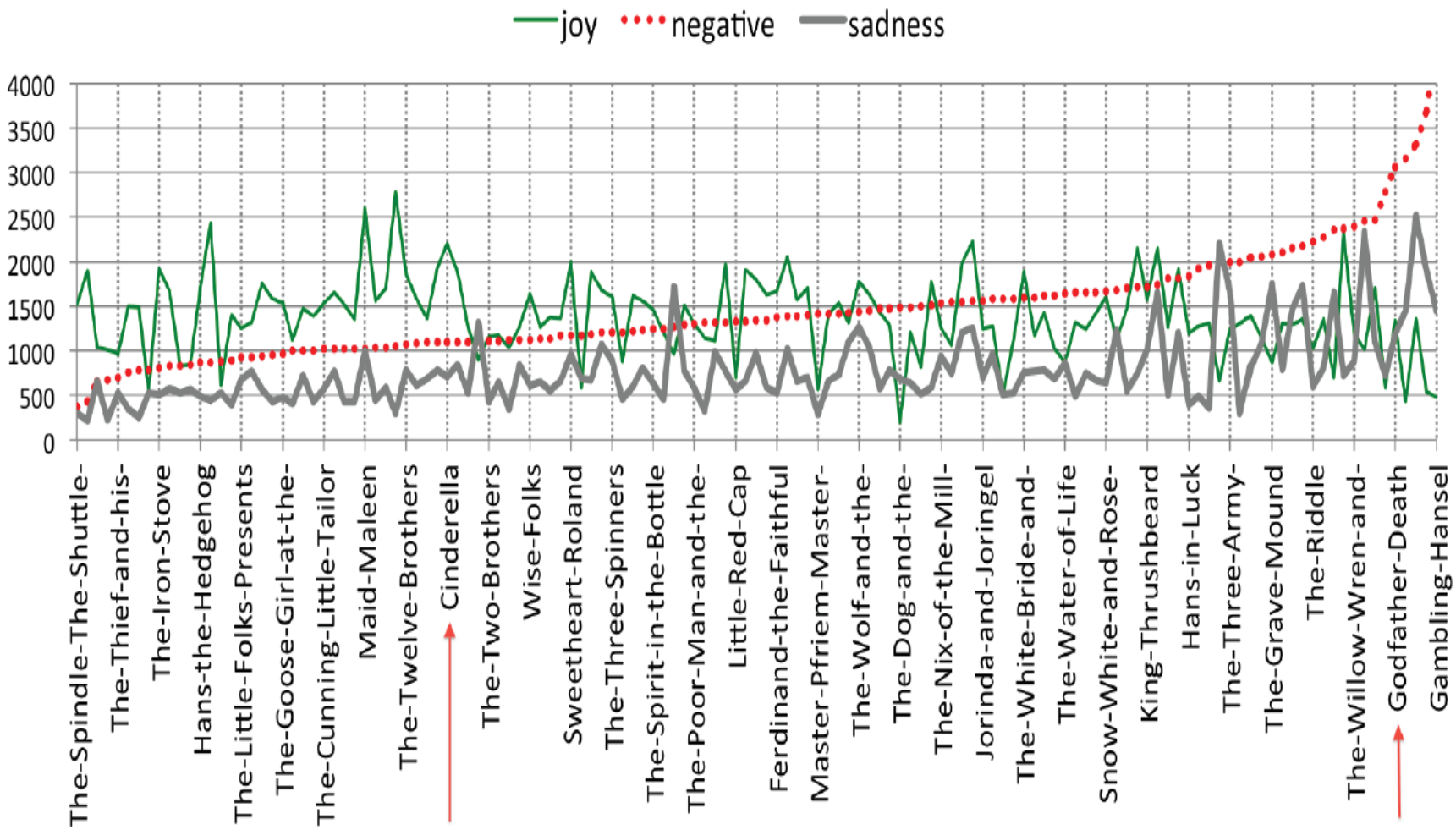


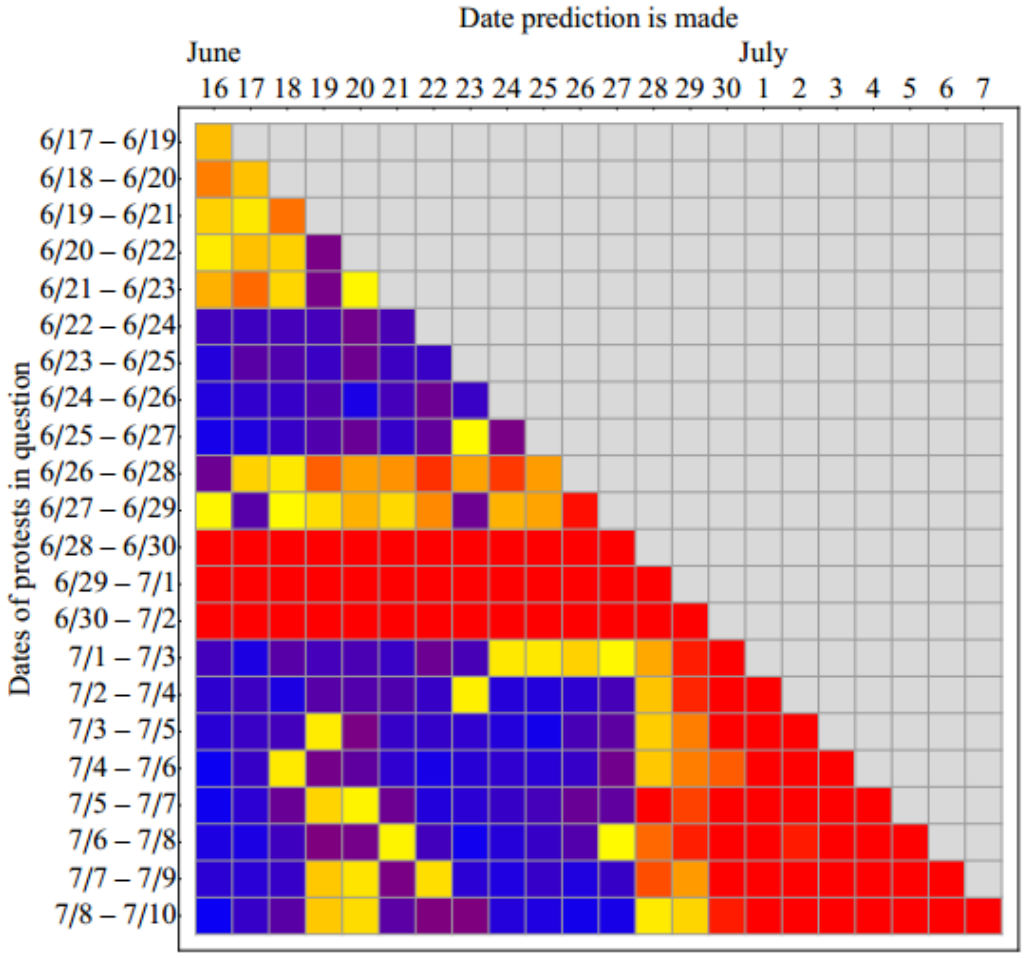
Figure 9: The Brothers Grimm fairy tales arranged in increasing order of negative word density (number of negative words in every 10,000 words). The plot is of 192 stories but the x-axis has labels for only a few due to lack of space. A user may select any two tales, say *Cinderella* and *Godfather Death* (follow arrows), to reveal Figure 10.

Representative Applications

- Mohammad and Yang (2011): tracked sentiments in emails see how genders differed on emotional axes. (e.g. women prefer words from the joy–sadness axis, whereas men prefer terms from the fear–trust axis.)

Representative Applications

- Based on public data collected from over 300,000 open content web sources in 7 languages, ranging from mainstream news to government publications to blogs and social media.
- Predictions of protests in Egypt around the time of the coup d'etat. (July 3, 2013)
- Yellow to red mark positive predictions and blue to purple negative, with redder colors indicating more positive votes.



https://en.wikipedia.org/wiki/2013_Egyptian_coup_d%27état

Predicting Crowd Behavior with Big Public Data

Nathan Kallus
Massachusetts Institute of Technology
77 Massachusetts Ave E40-149
Cambridge, MA 02139
kallus@mit.edu

ABSTRACT

With public information becoming widely accessible and shared on today's web, greater insights are possible into crowd actions by citizens and non-state actors such as large protests and cyber activism. We present efforts to predict the occurrence, specific timeframe, and location of such actions before they occur based on public data collected from over 300,000 open content web sources in 7 languages, from all over the world, ranging from mainstream news to government publications to blogs and social media. Using natural language processing, event information is extracted from content such as type of event, what entities are involved and in what role, sentiment and tone, and the occurrence time range of the event discussed. Statements made on Twitter about a future date from the time of posting prove particularly indicative. We consider in particular the case of the 2013 Egyptian coup d'état. The study validates and quantifies the common intuition that data on social media (beyond mainstream news sources) are able to predict major events.

sibility to public information on the web that future crowds may now be reading and reacting to or members of which are now posting on social media can offer glimpses into the formation of this crowd and the action it may take.

News from mainstream sources from all over the world can now be accessed online and about 500 million tweets are posted on Twitter each day with this rate growing steadily [14]. Blogs and online forums have become a common medium for public discourse and many government publications are offered for free online. We here investigate the potential of this publicly available information online for predicting mass actions that are so significant that they garner wide mainstream attention from around the world. Because these are events perpetrated by human actions, they are in a way endogenous to the system, enabling prediction.

But while all this information is in theory public and accessible and could lead to important insights, gathering it all and making sense of it is a formidable task. We here use data collected by Recorded Future (`www.recordedfuture.com`). Scanning over 300,000 different open content web sources in

3 Levels of Sentiment Analysis

- In general, sentiment analysis has been investigated mainly at three levels:
  - Document Level
  - Sentence Level
  - Entity & Aspect/Feature level

3 Levels of Sentiment Analysis

- **Document level**
  - The task at this level is to classify whether a whole document expresses a positive or negative sentiment.
    - For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product.
  - This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product)
  - Thus, it is not applicable to documents which evaluate or compare multiple entities.

3 Levels of Sentiment Analysis

- **Sentence level**
  - This level determines whether each sentence expressed a positive, negative, or neutral opinion.
    - Neutral usually means no opinion.
  - Related to subjectivity classification, which distinguishes objective/factual sentences from subjective sentences that express subjective views and opinions.
    - However, many objective sentences can also imply opinions:

      "We bought the car last month and the windshield wiper has fallen off. "
    - Conversely, many subjective sentences may not express any opinion or sentiment:

      "**I think** he went home after lunch."

3 Levels of Sentiment Analysis

- **Entity & Aspect/Feature level**
  - Both the document-level and sentence-level analyses do not discover what exactly people liked and did not like
  - Aspect level directly looks at the opinion itself, instead of documents, paragraphs, sentences, etc.
  - It is based on the idea that an opinion consists of a **sentiment** (positive or negative) and a **target** (of opinion).
  - Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better:

    "Although the service is not that great, I still love this restaurant."

    - This sentence is positive about the restaurant (emphasized), but negative about its service (not emphasized).

3 Levels of Sentiment Analysis

- **Entity & Aspect/Feature level**
  - In many applications, opinion targets are described by entities and/or their different **aspects**:

    "The iPhone's **call quality** is good, but its **battery life** is short."

    - This sentences evaluates two aspects: **call quality** and **battery life**, of iPhone (entity).
    - The sentiment on iPhone's call quality is positive, but the sentiment on its battery life is negative.
    - The call quality and battery life of iPhone are the opinion targets.

2 Types of Opinions

- 2 types of opinions need to be differentiated as well:
  - **Regular opinions**: express sentiments only on an particular entity or an aspect of the entity:

    "Coke tastes very good."

    - This sentence expresses a positive sentiment on the taste aspect of Coke.

  - **Comparative opinions**: compares multiple entities based on some of their shared aspects:

    "Coke tastes better than Pepsi."

    - This sentence compares Coke and Pepsi based on their taste aspect, and expresses a preference for Coke

Sentiment Lexicon

- The most important indicators of sentiments are **sentiment words**, also called opinion words.
  - These are words that are commonly used to express positive or negative sentiments.
    - Good, wonderful, and amazing are positive sentiment words
    - Bad, poor, and terrible are negative sentiment words.
  - Apart from individual words, there are also phrases and idioms:

    "It cost me an arm and a leg."
  - A list of such words and phrases is called a **sentiment lexicon** (or opinion lexicon).
  - Over the years, researchers have designed numerous algorithms to compile such lexicons. (**Sentiment Analysis and Opinion Mining**, Chapter 6; **Sentiment Analysis**, Chapter 7).

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages

WordNet

From Wikipedia, the free encyclopedia

WordNet is a lexical database for the English language.[1] It groups English words into sets of synonyms called *synsets*, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser,[2] its primary use is in automatic text analysis and artificial intelligence applications. The database and software tools have been released under a BSD style license and are freely available for download from the WordNet website. Both the lexicographic data (*lexicographer files*) and the compiler (called *grind*) for producing the distributed database are available.

Contents [hide]

A snapshot of WordNet's definition of itself.

- The SUMO ontology[44] has produced a mapping between all of the WordNet synsets (including nouns, verbs, adjectives and adverbs), and SUMO classes. The most recent addition of the mappings provides links to all of the more specific terms in the Mid-Level Ontology (MILO), which extends SUMO.
- OpenCyc,[45] an open ontology and knowledge base of everyday common sense knowledge, has 12,000 terms linked to WordNet synonym sets.
- DOLCE,[46] is the first module of the WonderWeb Foundational Ontologies Library (WFOL). This upper-ontology has been developed in light of rigorous ontological principles inspired by the philosophical tradition, with a clear orientation toward language and cognition. OntoWordNet[47] is the result of an experimental align WordNet's upper level with DOLCE. It is suggested that such alignment could lead to an "ontologically sweetened" WordNet, meant to be conceptually more rigorous, cognitively transparent, and efficiently exploitable in several applications.
- DBpedia,[48] a database of structured information, is linked to WordNet.
- The eXtended WordNet[49] is a project at the University of Texas at Dallas which aims to improve WordNet by semantically parsing the glosses, thus making the information contained in these definitions available for automatic knowledge processing systems. It is freely available under a license similar to WordNet's.
- The GCIDE project produced a dictionary by combining a public domain *Webster's Dictionary* from 1913 with some WordNet definitions and material provided by volunteers. It was released under the copyleft license GPL.
- ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images.[50] Currently, it has over 500 images per node on average.
- BioWordnet, a biomedical extension of wordnet was abandoned due to issues about stability over versions.[51]
- WikiTax2WordNet, a mapping between WordNet synsets and Wikipedia categories.[52]
- WordNet++, a resource including over millions of semantic edges harvested from Wikipedia and connecting pairs of WordNet synsets.[53]
- SentiWordNet, a resource for supporting opinion mining applications obtained by tagging all the WordNet 3.0 synsets according to their estimated degrees of positivity, negativity, and neutrality.[54]
- ColorDict, is an Android application to mobiles phones that use Wordnet database and others, like Wikipedia.
- UBY-LMF a database of 10 resources including WordNet.

SentiWordNet

SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. SentiWordNet is described in details in the papers:

SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining (view citations)

SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining (view citations)

The current version of SentiWordNet is 3.0, which is based on WordNet 3.0

License

Download

Download SentiWordnet 3.0

Other material

Micro-WordNet-Opinion 1.0 & 3.0

Micro-WordNet-Opinion is a dataset of human annotations of WordNet synsets that has been used to evaluate

SentiWordNet Interface

SentiWordNet can be imported like this:

```
>>> from nltk.corpus import sentiwordnet as swn
```

SentiSynsets

```
>>> breakdown = swn.senti_synset('breakdown.n.03')
>>> print(breakdown)
<breakdown.n.03: PosScore=0.0 NegScore=0.25>
>>> breakdown.pos_score()
0.0
>>> breakdown.neg_score()
0.25
>>> breakdown.obj_score()
0.75
```

Lookup

```
>>> list(swn.senti_synsets('slow')) # doctest: +NORMALIZE_WHITESPACE
[SentiSynset('decelerate.v.01'), SentiSynset('slow.v.02'),
SentiSynset('slow.v.03'), SentiSynset('slow.a.01'),
SentiSynset('slow.a.02'), SentiSynset('slow.a.04'),
SentiSynset('slowly.r.01'), SentiSynset('behind.r.03')]

>>> happy = swn.senti_synsets('happy', 'a')

>>> all = swn.all_senti_synsets()
```

Sentiment Lexicon

- Although sentiment lexicons are important or even necessary for sentiment analysis, they are not sufficient:
  - A positive or negative sentiment word may have opposite orientations in different application domains:

    "This camera **sucks**."

    "This vacuum cleaner really **sucks**."
  - A sentence containing sentiment words may not express any sentiment:

    "Can you tell me which Sony camera is **good** ?"

    "If I can find a **good** camera in this shop, I will buy it."

    - Both these sentences contain the sentiment word "good," but neither expresses a positive or negative opinion on any specific camera.

Sarcastic Sentences

- **Sarcastic sentences** with or without sentiment words are hard to deal with:

  "What a great car! It stopped working in two days."

  - Sarcasms are not so common in consumer reviews about products and services, but are very common in political discussions, which make political opinions hard to deal with.

- Many sentences without sentiment words can also imply opinions:

  "This washer uses a lot of water."

  - This sentence implies a negative sentiment about the washer since it uses a lot of resource (water).

  "After sleeping on the mattress for two days, a valley has formed in the middle."

  - This sentence expresses a negative opinion about the mattress.

Sentiment Analysis & NLP

- Sentiment analysis touches every aspect of **NLP**, such as coreference resolution, negation handling, and word sense disambiguation, which are **not** solved problems in NLP.
  - However, sentiment analysis is a highly restricted NLP problem because the system does not need to fully understand the semantics of each sentence or document but only needs to understand some aspects of it, i.e., positive or negative sentiments and their target entities.
  - We will use NLTK to serve our practical NLP needs.

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

Some simple things you can do with NLTK

SEARCH

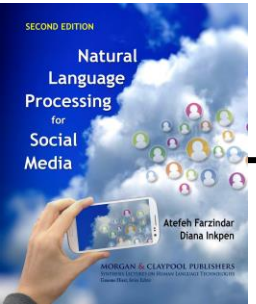
Go

Challenges In Social Media Texts (Optional)

Challenges In Social Media Texts

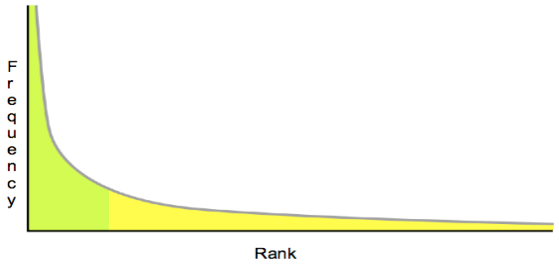
- The use of social networks has made everybody a potential author, so the language is now **closer to the user** than to any prescribed norms.

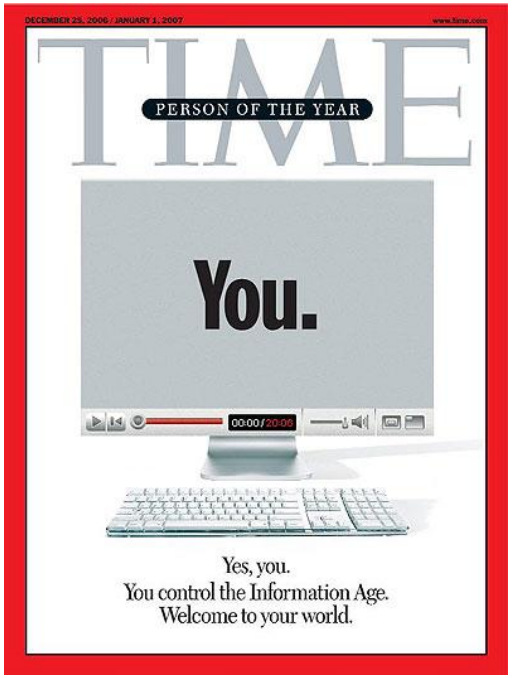
- Blogs, tweets, posts and status updates are written in an **informal**, **conversational** tone (a "stream of consciousness" rather than the meticulously edited work that might be expected in traditional print media.)



Characteristics of Social Media (Revisited)

- **Consumers become Producers**
- Rich User Interaction
- **User-Generated Contents**
- Collaborative Environment
- Collective Intelligence
- Long Tail





Broadcast Media
Filter, then Publish

Social Media
Publish, then Filter

Challenges In Social Media Texts

- Several characteristics of social media text are of particular concern.
  - For example, the 140-character limit imposed on Twitter posts makes for individual tweets that are rather contextually impoverished compared to more traditional documents.
  - Redundancy can become a problem over multiple tweets, due in part to the practice of retweeting posts.
  - **feature sparseness** (next slide)
- The particularities of a given medium and the way in which that medium is used can have a profound effect on text mining.

Feature Sparseness

- Tweets are short
  - No more than 140 characters, usually no more than ten words or so.
- This means that Twitter posts are **feature sparse**, and hence comparisons between posts is difficult.

  → big problem for text clustering, which is highly sensitive to the features chosen for comparison.
- This problem could be alleviated somehow by
  - Expanding terms in the tweets to include relevant similar terms (essentially adding additional features)
  - Selecting more descriptive features (e.g. just named entities)
  - Using n-grams instead of unigrams
  - Some combination of the three.

Challenges In Social Media Texts

- Standard text mining/NLP methods applied to social media texts are also confronted with difficulties:
  - **Inconsistent (or absent) punctuation and capitalization** can make detection of sentence boundaries quite difficult - sometimes even for human readers:

    #qcpoli enjoyed a hearty laugh today with #plq debate audience for @jflisee #notrehome tune was that the intended reaction?

  - **Emoticons**, incorrect or non-standard spelling, and rampant abbreviations complicate tokenization and part-of-speech tagging, among other tasks. (→ **Noise**, see next slide)

Noise

- While most standard NLP techniques were developed for long, structured, grammatical text, tweets are short, colloquial, and ungrammatical. (See more examples on the following slides.)
  - Tweets also frequently contain other non-standard tokens, including **emoticons**
  - Users frequently **misspell words** either unintentionally (teh, waht) or intentionally, by expanding words, abbreviating words, or using lexical/numeric substitutions:
    - E.g. heyyyyyy, gooooooood, loooovveeeee, rly, c u l8r,…
  - **Acronyms** (lol, smh), hashtags, user mentions, or Twitter specfic terminology indicating re-tweeted posts (RT) and trending topics (TT).

Grammaticality

- **Grammaticality**, or frequent lack thereof, is another concern for any syntactic analyses of social media texts
  - Fragments can be as commonplace as actual full sentences
  - The choice between “there,” “they are,” “they’re,” and “their” can seem to be made at random…

TABLE I
SAMPLE TWEETS

Never say never.....dont let me goo dont let mee gooo dont let me gooooo....
@user13431 when r u commin to Montreal
#bestfeeling is feeling like u mean the world to someone
My work buddy 'go smoke' like 3 times already
mai8mai RT @user1341 : Support Breast Cancer Awareness. Add A #twibbon To Your Avatar Now!!
I'm so #overyou Didn't even know it was possible!!!

Normalization

- As a result, two tweets with alternate spellings of some word may not be considered related, when in fact they are.
- **Normalization** of tweets remains a difficult problem, but fortunately some progress has been made, and we will discuss normalizing text, in general, and normalizing tweets, in particular, later in this course. (See some examples on the next slide)

Original: @user3419 nay lol y u say dat?&wat u doing 2day?

Post-normalization: No, why did you say that? What you doing today?

Original: 1001 colors: Contemporary art from Iran <URL> #Iran #culture #Art

Post-normalization: 1001 colors: contemporary art from Iran <URL>.

Original: it's soo quiet, it's like I'm goin die

Post-normalization: It is so quiet, it is like i am going to die.

Original: #worstfeeling buyin a fresh laptop..then ur screen blowz out :((

Post-normalization: worst feeling is buying a fresh laptop.. then your screen blowz out.

---

*Original:*This is superb Grape+apple splash with manggo juice, super!

Post-normalization: This is superb grape + Apple splash with mango juice, Super!

Original: @user31903 u n ur fam can n if u interested ill b n touch w u bout it

Post-normalization: You and your family can and if you interested Ill be and touch with you about it.

Original: RT: @user4191 BEAUTIFUL CREATURES has a new #website designed by @user4192!

Post-normalization: Beautiful creatures has a new website designed by @user4192!

---

researchers studying SMS normalization have chosen to evaluate their results with the BLEU metric, it might not be the best choice. The BLEU scoring metric was designed for evaluating translations from one language to another, not for evaluating the results of noisy text normalization. Because of this, a better BLEU score does not necessarily mean a better translation. For example, the subjectivity of the human annotators could cause substantial variation in BLEU scores. In papers such as [8] their corpora was only annotated by two people. The fact that there were 10 annotators could have lead to inconsistencies in the scoring data. For example, although annotators were instructed to expand contractions, some annotators chose to translate *Im* as *I'm*, instead of *I am*. BLEU scores are obtained by comparing the similarities between *n-grams* of the hypothesized translation and gold standards, so errors such as this could have detrimental effects on the score, despite the fact that "I'm" and "I am" are grammatically equivalent.

Even if we ignore the issue of the applicability of BLEU as an evaluator itself, there are still several problems with the BLEU metric itself. The relationship between BLEU scores and human judgment is questionable. Papers such as [16] have suggested that an increase in BLEU score may not correlate with an increase in translation quality. In fact, on a test of several machine translation systems, the correlation between human and BLEU scores was found to be as low as .38 in some cases. One example where the BLEU score performed poorly was on the tweet *@user12493 I'm following u now should I hold on tight?*. The translation generated by the normalization system was *I'm following you now, should I hold on tight*". This seems like a perfectly acceptable translation. However, the human annotator translated the tweet as *I'm following you. Now, should I hold on tight?*. BLEU scores this translation at .43. However, both translations are acceptable.

Joseph Kaufmann and Jugal Kalita. 2010. Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India

Syntactic Normalization of Twitter Messages

Max Kaufmann

Abstract—**The use of computer mediated communication such as emailing, microblogs, Short Messaging System (SMS), and chat rooms has created corpora which contain incredibly noisy text. Tweets, messages sent by users on Twitter.com, are an especially noisy form of communication. Twitter.com contains billions of these tweets, but in their current state they contain so much noise that it is difficult to extract useful information. Tweets often contain highly irregular syntax and nonstandard use of English. This paper describes a novel system which normalizes these Twitter posts, converting them into a more standard form of English, so that standard machine translation (MT) and natural language processing (NLP) techniques can be more easily applied to them. In order to normalize Twitter tweets, we take a two step approach. We first preprocess tweets to remove as much noise as possible and then feed them into a machine translation model to convert them into standard English. Together, these two steps allow us to achieve improvement in BLEU scores comporable to the improvements achieved by SMS normalization**

novel words, and interjections. A word may be written using a phonetic spelling (*nite* instead of *night*), or combined with other frequently used words into an acronym (*omg* instead of *oh my god*). Twitter users also have little regard for the proper use of capitalization and punctuation. Capitalization in a tweet may signal a proper noun or a sentence boundary, but it may also be used for something as arbitrary as emphasizing a certain segment. Punctuation may signal sentence boundaries, but it might also be used to create an emoticon. There are some deviations that are standard and systematic, but new variations can be created at any time, making the process of modeling the language extremely difficult. Additionally, Twitter users frequently use symbols to encode meta-content, such as who the tweet was directed to, or the topics to which it pertains. This meta-content sometimes is integrated into the syntax of the tweet, but there is no guarantee that it will be. In

Other Forms of Noise

- Social media are also much noisier than traditional print media.
  - Like much else on the Internet, social networks are plagued with **spam**, **ads**, and all manner of other unsolicited, irrelevant, or distracting content.
  - Even by ignoring these forms of noise, much of the genuine, legitimate content on social media can be seen as irrelevant with respect to most information needs.
- In a study, the authors collected over 40,000 rating of tweets from follows. Only 36% of them tweets were rated as "worth reading."
  - The least valued tweets were so-called presence maintenance posts (e.g., "Hullo twitter!").
  - Pre-processing to filter out spam and other irrelevant content, or models that are better capable of coping with noise, are essential in mining social media texts.

Other Forms of Noise

- A major challenge facing **event detection** from tweets is therefore to separate the mundane and polluted information from interesting real-world events.
  - In practice, highly scalable and efficient approaches are required for handling and processing the increasingly large amount of Twitter data (especially for real-time event detection).
  - In addition, different events may enjoy different popularity among users, and can differ significantly in content, number of messages and participants, time periods, inherent structure, and causal relationships.

Subjectivity

- Across all forms of social media, **subjectivity** is an ever-present trait
  - While traditional news texts may strive to present an objective, neutral account of factual information, social media texts are much more subjective and opinion-laden.
  - Whether or not the ultimate information need lies directly in sentiment analysis, or in factual information, subjective information plays a much greater role in semantic analysis of social texts.

    → Subjectivity Analysis/Classification

Topic Drift

- **Topic drift** is also much more prominent in social media than in other texts, both because of the conversational tone of social texts and the continuously streaming nature of social media. (See an example on the next slide)

- Citation

@Inbook{Fei2015,author="Fei, Yueand Hong, Yihongand Yang, Jianwu",editor="Hanbury, Allanand Kazai, Gabriellaand Rauber, Andreasand Fuhr, Norbert",chapter="**Handling Topic Drift for Topic Tracking in Microblogs**",title="Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings",year="2015",publisher="Springer International Publishing",address="Cham",pages="477--488",isbn="978-3-319-16354-3",doi="10.1007/978-3-319-16354-3_52",url="http://dx.doi.org/10.1007/978-3-319-16354-3_52"}

Handling Topic Drift for Topic Tracking in Microblogs

Yue Fei, Yihong Hong, and Jianwu Yang*

Institute of Computer Science and Technology Peking University, China
{feiyue,hongyihong,yangjw}@pku.edu.cn

Abstract. Microblogs such as Twitter have become an increasingly popular source of real-time information, where users may demand tracking the development of the topics they are interested in. We approach the problem by adapting an effective classifier based on Binomial Logistic Regression, which has shown to be state-of-art in traditional news filtering. In our adaptation, we utilize the link information to enrich tweets' content and the social symbols to help estimate tweets' quality. Moreover, we find that topics are very likely to drift in microblogs as a result of the information redundancy and topic divergence of tweets. To handle the topic drift over time, we adopt a cluster-based subtopic detection algorithm to help identify whether drift occurs and the detected subtopic is regarded as the current focus of the general topic to adjust topic drift. Experimental results on the corpus of TREC2012 Microblog Track show that our approach achieves remarkable performance in both T11SU and F-0.5 metrics.

Topic Drift: An Example

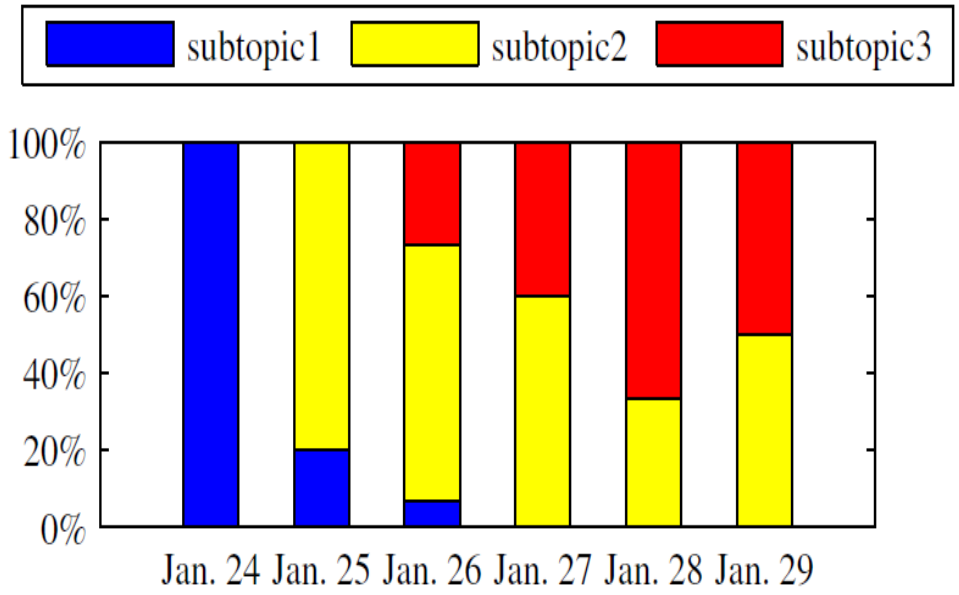


Fig. 1. Distribution of each subtopic's tweet percentage over time of Topic "BBC World Service staff cuts"

Allan Hanbury Gabriella Kazai
Andreas Rauber Norbert Fuhr (Eds.)

LNCS 9022

Advances in Information Retrieval

37th European Conference on IR Research, ECIR 2015
Vienna, Austria, March 29 – April 2, 2015
Proceedings

Springer

Proceedings of RIAO '94 Conference 1994

Oct 11, 1994 (presentation published)

(9)

2 copies text color 8/10/95

Document Retrieval Using Linguistic Knowledge

Elizabeth D. Liddy, Woojin Paik, Edmund S. Yu, Mary McKenna
Syracuse University
4-206 Center for Science & Technology
Syracuse, New York 13244-4100
liddy@mailbox.syr.edu

Abstract

The theoretical goal underlying the DR-LINK Text Retrieval System is to represent and match documents and queries at the various linguistic levels at which human language conveys meaning. Accordingly, we have developed a modular system which processes and represents text at the lexical, syntactic, semantic, and discourse levels of language. In concert, these levels of processing permit DR-LINK to achieve a level of intelligent retrieval beyond more traditional approaches. In addition, the rich annotations to text produced by DR-LINK are replete with much of the semantics necessary for more precise information extraction tasks.

The Text Structurer is based on discourse linguistic theory which suggests that texts of a particular type have a predictable text-level structure which serves as an indication of how and where certain information endemic to a text-type will be conveyed. We have implemented a Text Structurer for the newspaper text-type, which produces an annotated version of a news article in which each clause or sentence is tagged for the specific slot it instantiates in the news-text model, e.g. MAIN EVENT, EXPECTATION, CONSEQUENCE. The structural annotations are used to respond more precisely to information needs expressed in queries, where some aspects of relevancy requirements such as time, source, intentionality, and state of completion can only be met by understanding a query's discourse requirements (e.g. the **consequences** of automation; a **proposed** theme park development).

The Text Structurer assigns news-text component labels to document clauses/sentences on the basis of four types of linguistic evidence learned from text. We have reduced the matching complexity via a function that maps the thirty-eight news-text components which are recognized in documents to seven meta-component requirements which are recognized in queries. This allows the system to impose fine-level structure on newspaper articles with excellent precision and to map this fuller set of text components to the appropriate level of discourse requirement specificity typically expressed in queries.

The Subject Field Coder (SFCoder) uses an established semantic coding scheme from the machine-readable Longmans Dictionary of Contemporary English (LDOCE) to tag each word in a text with its disambiguated subject code (e.g. Agriculture, Military, Political Science) and to then produce a fixed-length, subject-based vector representation of each document's and query's contents. Using the SFCoder, each text or sub-text is represented as a vector of the

Challenges In Social Media Texts

- The dynamic nature of social media can be seen as an additional source of complexity that may hamper traditional summarization approaches.
- It is also an opportunity, making available additional context that can aid in **summarization** or making possible entirely new forms of summarization.
  - Hu et al. [2007a] suggest summarizing a blog post by extracting representative sentences using information from user comments.
  - Chua and Asur [2012] exploit temporal correlation in a stream of tweets to extract relevant tweets for event summarization.
  - Lin et al. [2009] address summarization not of the content of posts or messages, but of the social network itself by extracting temporally representative users, actions, and concepts in Flickr data.

Comments-Oriented Blog Summarization by Sentence Extraction*

Summarizing a blog by extracting representative sentences from comments.

Meishan Hu, Aixin Sun and Ee-Peng Lim
Centre for Advanced Information Systems
School of Computer Engineering
Nanyang Technological University, Singapore
{hu0004an,axsun,aseplim}@ntu.edu.sg

ABSTRACT

Much existing research on blogs focused on posts only, ignoring their comments. Our user study conducted on summarizing blog posts, however, showed that reading comments does change one's understanding about blog posts. In this research, we aim to extract representative sentences from a blog post that best represent the topics discussed among its comments. The proposed solution first derives representative words from comments and then selects sentences containing representative words. The representativeness of words is measured using *ReQuT* (i.e., *Re*ader, *Qu*otation, and *T*opic). Evaluated on human labeled sentences, *ReQuT* together with summation-based sentence selection showed promising results.

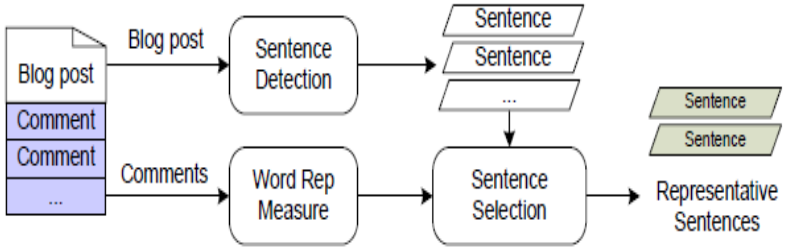


Figure 1: Comments-oriented blog summarization

conducted a user study on summarizing blog posts by labeling representative sentences in those posts. Significant differences between the sentences labeled before and after reading comments were observed.

In this research, we therefore focus on the problem of comments-oriented blog post summarization. The task is

Automatic Summarization of Events from Social Media

Freddy Chong Tat Chua*
Living Analytics Research Centre
Singapore Management University
80 Stamford Road, Singapore
freddy.chua.2009@smu.edu.sg

Sitaram Asur
Social Computing Lab
visited Hewlett Packard Research Labs
Palo Alto, California, USA
sitaram.asur@hp.com

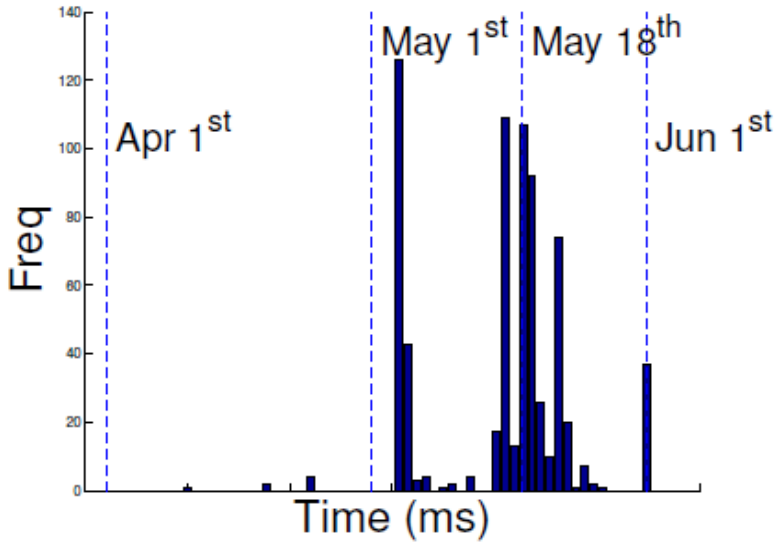
Temporal correlation of tweets can be used for for event summarization.

ABSTRACT

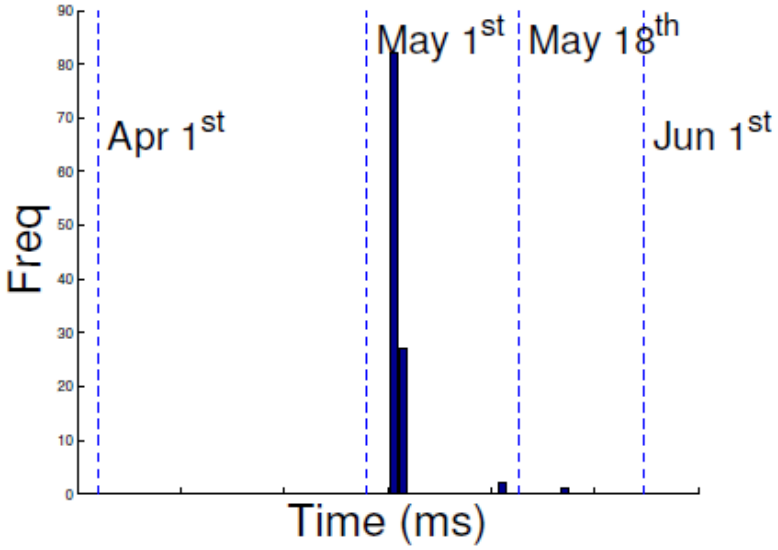
Social media services such as Twitter generate phenomenal volume of content for most real-world events on a daily basis. Digging through the noise and redundancy to understand the important aspects of the content is a very challenging task. We propose a search and summarization framework to extract relevant representative tweets from an *unfiltered tweet stream* in order to generate a coherent and concise summary of an event. We introduce two topic models that take advantage of temporal correlation in the data to extract relevant tweets for summarization. The summarization framework has been evaluated using Twitter data on four real-world events. Evaluations are performed using Wikipedia articles on the events as well as using Amazon Mechanical Turk (MTurk) with human readers (MTurkers). Both experiments show that the proposed models outperform traditional LDA and lead to informative summaries.

a flood of information propagated through these networks. By closely monitoring these streams of information, prior research have shown that it is possible to detect real world events from Twitter [21, 23, 24, 29, 30]. An event refers to any concept of interest that gains the attention of the populace. Examples of real-world events range from global catastrophes such as earthquakes [23], political protests or unrest [30], to launches of new consumer products.

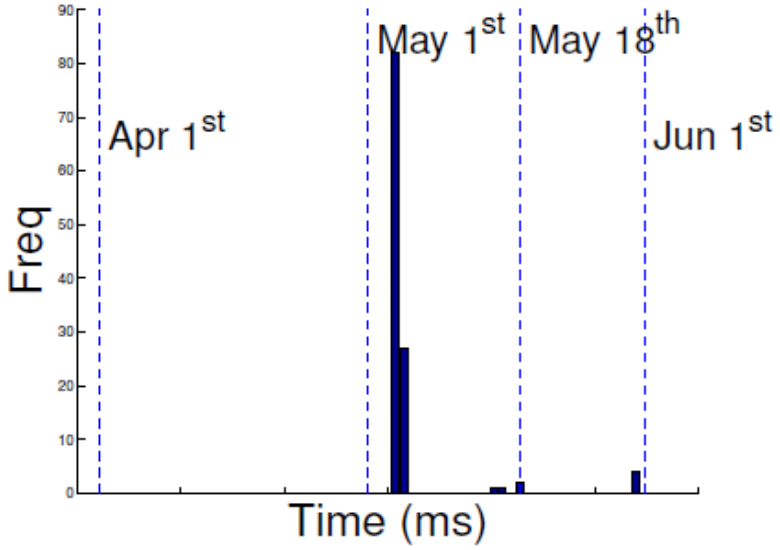
The easiest way to extract tweets related to an event is through a search query. However, for popular events, this typically results in a significantly large stream of tweets, which makes the task of understanding the aspects of the event and the opinion of people, a difficult and mostly futile task. It has been observed that, despite the high frequency, the actual information content in the tweet stream is fairly limited [7, 25]. This is due to the fact that several of the tweets contain redundant information. Also, many of the tweets that are returned by a search query are not relevant to the event. This is due to ambiguity in the search keywords



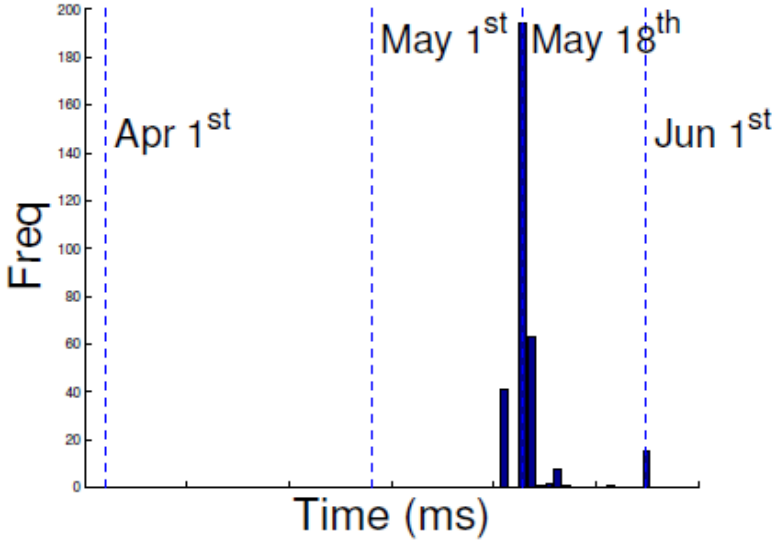
(a) Frequency of "price"



(b) Frequency of "$28"



(c) Frequency of "$35"



(d) Frequency of "$38"

New Forms of Summarization

- Lin et al. [2009] address summarization not of the content of posts or messages, but of the social network itself by extracting temporally representative users, actions, and concepts in Flickr data.

- Citation:

  Hui Lin, Jeff Bilmes, and Shasha Xie. Graph-based submodular selection for extractive summarization. In *the eleventh biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2009)*, pages 381–386. IEEE, 2009. DOI: 10.1109/ASRU.2009.5373486.10

Challenges In Social Media Texts

- Network analysis - exploring additional structure in the text corpus