ACADEMIC YEAR 2026/27

Text Mining & Search [DS] [CS] / Information Retrieval

Course Overview & Introduction

TEXT MINING & SEARCH

Prof. Marco Viviani

University of Milano-Bicocca

Department of Informatics, Systems and Communication

INFORMATION RETRIEVAL

Dr. Georgios Peikos

University of Milano-Bicocca

Department of Informatics, Systems and Communication

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Ikr3 Lab — information and knowledge representation, retrieval, and reasoning LAB

Course Overview – Why the two courses begin together?

Text Mining & Search and Information Retrieval share the same foundations.

- Both courses work with text. We begin with the methods used to process, analyse, and represent documents.
- Both courses study search (i.e. Information Retrieval). You will learn how a system finds relevant documents, given a user's information need.
- Shared lessons will take place till mid-October
- The courses then develop different areas of emphasis.
  - Data Science continues with text mining tasks, applications, and search.
  - Computer Science continues with Information Retrieval focusing on the theoretical aspects, technical design, and evaluation.

Shared Part I – Overview [Tentative]

01 **Introduction** — Overview of the course and introduction to Text Mining and Tasks

02 **Basic Text Processing** — Preparing and normalizing raw textual data for computational analysis

03 **Text Representation** — Transforming textual content into structured and numerical representations

04 **Part-of-Speech Tagging & NER** — Identifying grammatical categories and relevant entities in text

05 **Data Structures for Textual Search** — Organizing and storing textual information for efficient computational access

06 **Word and Contextual Embeddings** — Representing words and their contexts in continuous vector spaces

Text Mining & Search Part II – Overview [Tentative]

01 **Text Classification** — In depth analysis of the task's inner workings

02 **Text Clustering** — In depth analysis of the task's inner workings

03 **Topic Modeling** — In depth analysis of the task's inner workings

04 **Text Summarization** — In depth analysis of the task's inner workings

05 **Information Retrieval** — Introduction, Web Search, Contextual Search, Personalized Search

06 **Introduction to RAG** — An overview from Naive to Advanced and Modular RAG systems

Information Retrieval Part II – Overview [Tentative]

01 **Introduction to Search Engines** — Information needs, queries, documents, relevance, and search applications

02 **Indexing and Retrieval Structures** — Inverted indexes, posting lists, and efficient retrieval pipelines

03 **Relevance Estimation and Classical IR Models** — Boolean model, vector-space models, probabilistic models, and BM25

04 **Neural Information Retrieval** — Dense, sparse, hybrid, and multi-stage neural retrieval models

05 **Evaluation of IR Systems** — Test collections, relevance judgments, evaluation metrics, and significance testing

06 **IR in Conversational AI** — Query rewriting, reranking, Retrieval-Augmented Generation (RAG), and citation support

07 **Advanced Topics** — Domain-specific search, learning to rank, agentic search, and emerging research directions

Course and Exams – Overview

01 Course Format

Learning Process

- **40 Hours** of lectures
  - Practical cases & exercises
- **12 Hours** of hands-on laboratories
  - Use of open-source software
  - Delivered by PhD students with expertise in the topics described above

02 Evaluation (Exams)

Testing & Core Assessment

- **Written Examination** (30/30)
  Theoretical questions & practical exercises designed to evaluate your ability to architect a search engine for a given use case.
- **Laboratory Project** (Team Project) (+3 points)
  Each team will receive a dataset and a real-world search problem.

  Each team will design, implement, and evaluate its own search engine using open-source software.

Course and Exams – Additional Details

- Written Exam. You must write at least 18/30, so we can add to that the project additional points.
- Example Evaluation. If your written exam score is 23/30, and you did a very good project that awarded your team +3 points, your final score will be 26/30.
- Honors ("Lode"). Assigned with a total score of at least 32/30.
- Best project in Information Retrieval course. It will be awarded 4 points.

Course and Exams – Material and Communication

- Slides and Reading Material. Will be uploaded in the dedicated course pages in the e-learning platform.
- Communication. You can write an email to georgios.peikos@unimib.it and to marco.viviani@unimib.it; make sure to share from which course you are.
- Updates. All communications from us will be received via the e-learning platform and the announcements their, so make sure that you subscribed to the courses.

Questions?

Feel free to ask anything about the course, exams, projects.

Get to know each other – Raise Hand

Topic	Beginner	Intermediate	Advanced
Python and working with data	**New to me** e.g., Never written Python or used pandas	**I know the basics** e.g., Loops, lists, basic numpy/pandas	**I have used it in a project** e.g., Data cleaning, scraping, or analysis scripts
Vectors and probability	**New to me** e.g., Unfamiliar with dot products or Bayes' rule	**I have studied them** e.g., Matrix operations, **conditional probability** in class	**I can apply them** e.g., Computing cosine similarity or probability distributions in code
Machine learning and evaluation	**New to me** e.g., Haven't used scikit-learn or ML models	**I know the concepts** e.g., Precision/Recall, Overfitting, **Un/Supervised learning**	**I have trained and tested a model** e.g., Trained a classifier with train/test split & evaluated F1-score
Text processing or search systems	**New to me** e.g., Haven't built search engine or processed text	**I know the concepts** e.g., **Tokenization**, TF-IDF, Stemming, Inverted index	**I have built something with them** e.g., Built an inverted index, Lucene app, or NLP pipeline

Introduction to Text Mining & Search

Text Mining and Search – Definitions and Term Clarification (1/2)

What is text mining?

Text **(data)** mining is the discovery of previously **unknown information or patterns** through the analysis of text collections [1,Hearst, 1999].

*"Text mining is a burgeoning new field that attempts to glean **meaningful information** from natural language text. It may be loosely characterized as the **process** of analyzing text to **extract information** that is useful for **particular purposes**."* [2, Witten, 2005]

[1] Marti A. Hearst. 1999. Untangling Text Data Mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 3–10, College Park, Maryland, USA. Association for Computational Linguistics.
[2] Witten, I. H. (2005). Text mining. In M. P. Singh (Ed.), Practical Handbook of Internet Computing (pp. 14-1–14-22). Chapman & Hall/CRC.

Text Mining and Search – Definitions and Term Clarification (2/2)

What is Information Retrieval (search)?

Information Retrieval (IR) is a field concerned with the **representation**, **storage**, **organization**, and **access of information** items so that users can locate those relevant to their needs.

Given a user's request, IR systems identify and rank the information that is most relevant to the user's information need.

[1] Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to information retrieval. Vol. 39. Cambridge: Cambridge University Press, 2008.

Given a collection of textual information,

Text mining

Detect patterns/information in articles and track their coverage over time.

For instance, analyze how a topic involved over time.

Search

Return the articles most relevant to **“solar energy in Milan”**.

Text Mining
The History

A historic Overview of Text Mining – The 1990s

Discovering knowledge across documents

- Early research extended knowledge discovery from structured databases to unstructured **textual collections of documents**.
- Researchers **analysed concepts across documents** to identify **patterns** and **relationships** across them.
- Overlapping definitions of data mining and text mining.

Source: Feldman, Ronen, and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)." In KDD, vol. 95, pp. 112-117. 1995.

Text Mining and Data Mining – Debate Based on early definitions (1990s)

Data Mining

Implicit & Hidden Information

Data mining is characterized as the extraction of **implicit**, previously **unknown**, and potentially useful information from data [Witten and Frank, 2000].

Key Characteristics:

- Information is **implicit** within input data.
- Data is hidden and unknown prior to mining.
- Requires automated techniques to extract.

Text Mining

Explicit & Stated Information

In text mining, the information to be extracted is **clearly** and **explicitly** stated directly within the text. It is not hidden at all.

Key Characteristics:

- Information is **explicitly** present in text.
- Focuses on converting text into machine-ready formats.
- Directly processed without needing human intermediaries.

A historic Overview of Text Mining – The early 2000s

Extracting useful information

- Broader definitions emphasized **extracting information useful** for a particular purpose.
- **Text mining** included identifying names, extracting keyphrases, and assigning document categories.
- Under this interpretation, the extracted information could **already appear explicitly** in the text.

A historic Overview of Text Mining – The late 2000s to 2010s

A family of text analysis tasks

- Text mining increasingly served as an **umbrella** term for **classification**, **clustering**, **extraction**, and other **analysis tasks**.
- Development of systems that combined text processing, statistical methods, machine learning, and human interpretation.
- Applications included topic analysis, sentiment analysis, and summarization to support decisions.

A historic Overview of Text Mining – Late 2010s to today

Neural and generative methods (Large Language Models)

- Pretrained language models provide **contextual representations of text**.
- Generative models can extract structured information, and generate summaries, among other tasks.
- These novel approaches changed how we perform text mining, i.e. how we develop systems for TM.
- The objective remains obtaining useful information and patterns from text.

Text Mining
Around Us

Text Mining – Texts we produce in our everyday lives

Digital & Media Content

- Emails
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- User generated content in Social media
  (course on Social Media Analytics)

Enterprise & Specialized

- Customer complaint letters
- Contracts
- Transcripts of phone calls with customers
- Technical documents
- Scientific papers
- Health related information

Bodo's Bagels

186 reviews Details

$ · Bagels, Breakfast & Brunch, Sandwiches Edit

Write a Review | Add Photo | Share | Bookmark



1418 Emmet St N
Charlottesville, VA 22903
Edit
Get Directions
(434) 977-9598
Message the business
bodosbagels.com



See all 36



"Almost any combination of bagel, cream cheese or spread or sandwich you could dream of you can find at Bodos." in 38 reviews
$0.60 Cream Cheese



"A few favorite items would include the Everything bagel with the Deli Egg which has a tasty meaty center encased in steaming hot eggs." in 4 reviews



"There's a reason why Bobo's has been in business since well before I was a UVa." in 10 reviews

Today 6:30 am - 8:00 pm Closed now
Full menu
$$$$ Price range Under $10

Hours

Mon	6:30 am - 8:00 pm
Tue	6:30 am - 8:00 pm Closed now
Wed	6:30 am - 8:00 pm
Thu	6:30 am - 8:00 pm
Fri	6:30 am - 8:00 pm
Sat	7:00 am - 8:00 pm
Sun	8:00 am - 4:00 pm

Edit business info

Recommended Reviews

Search reviews

Sort by Highest Rated

English (186)

Hilton Times Square

4,919 Reviews | #70 of 467 Hotels in New York City | Certificate of Excellence

+1 855-271-3621 | Hotel deals | Hotel website | 234 West 42nd Street, New York City, NY 10036

Special Offer TripAdvisor Special Offer

PriceFinder

Enter dates for best prices

Check In | Check Out

Check Availability

Book on tripadvisor

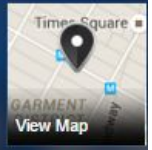
or compare prices from up to 200 sites including:

Booking.com | travelocity | Expedia









★★★★☆ | Pets Allowed | Luxury | Times Square / Theater District

Overview | Reviews (4,919) | Photos (1,654) | Location | Amenities | Q&A (129) | Room Tips (1,085) | Save

4,919 Reviews from our TripAdvisor Community

Write a Review | Add Photo

Text Mining – In everyday life

Messages

Spam filtering, suggested replies, and grouping similar conversations.

Reviews and feedback

Sentiment and recurring complaints across many comments.

News and reading

Topic labels, summaries, and links between related stories.

Documents

Extracting names, dates, or other fields from written material.

What Data/Computer Scientists (you) build with text

Organize

For instance, route support tickets or classify documents by topic.

Extract

Identify entities and relations in articles, reports, or scientific papers.

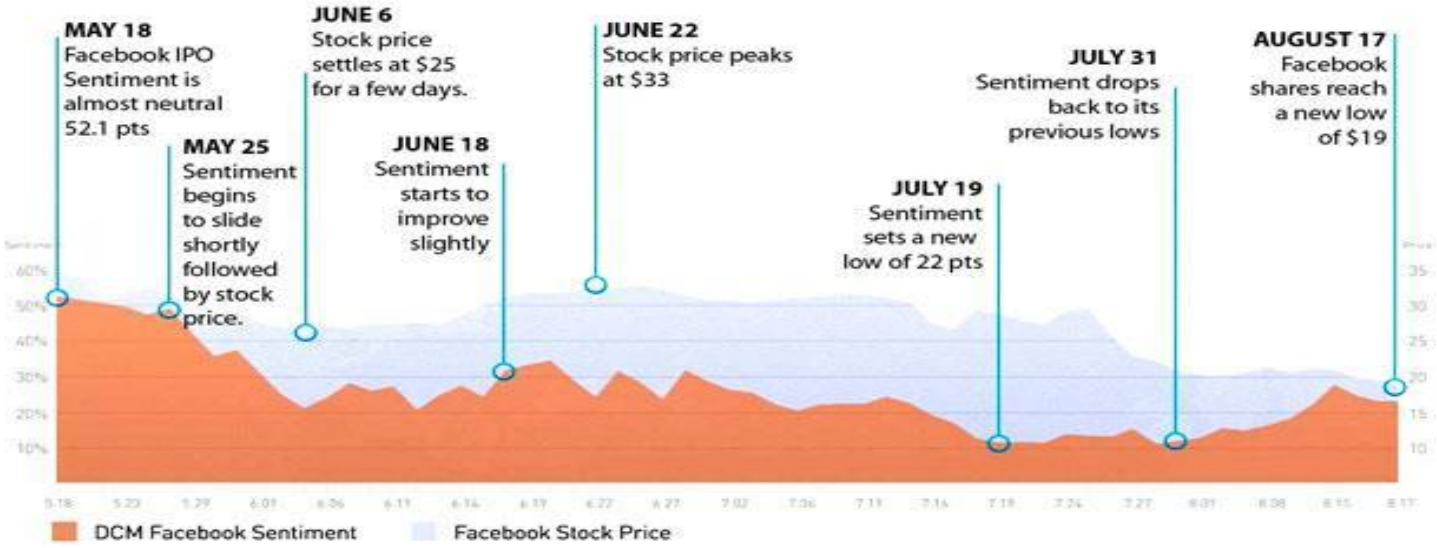
Explore

Group documents and detect recurring themes in large collections.

Answer

Find evidence in a corpus and use it to support an answer.

Text Mining – From the perspective of a Data/Computer Scientist



Text Mining
Main Tasks

Text Mining Tasks | Text Classification

SUPERVISED LEARNING

Text Classification

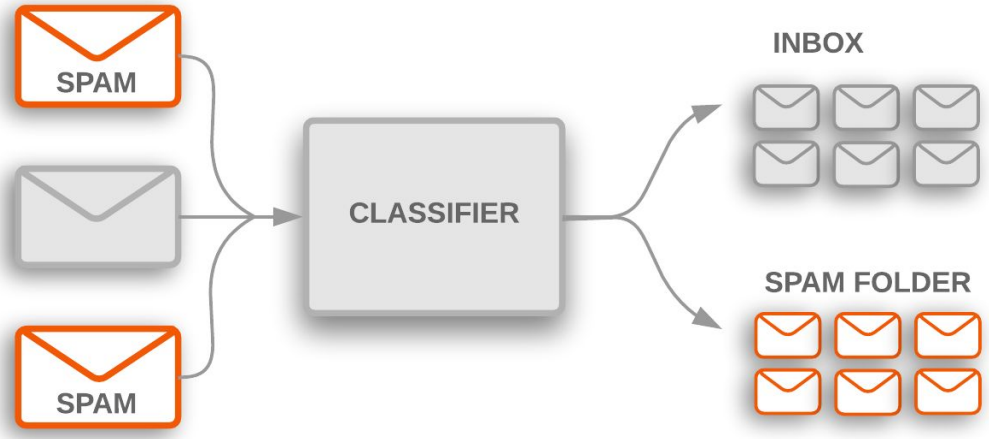
The assignment of natural language documents to predefined categories according to their content. **[Sebastiani, 2002]**

Predefined Categories

Documents are mapped into **existing, target classes** based on textual analysis and supervised model rules.

Key Applications

Variety of uses across domains, including Sentiment Analysis, Topic Labeling, and Automated Document Routing.


SPAM
SPAM
CLASSIFIER
INBOX
SPAM FOLDER


NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Text Mining Tasks | Text Clustering

UNSUPERVISED LEARNING

Text Clustering

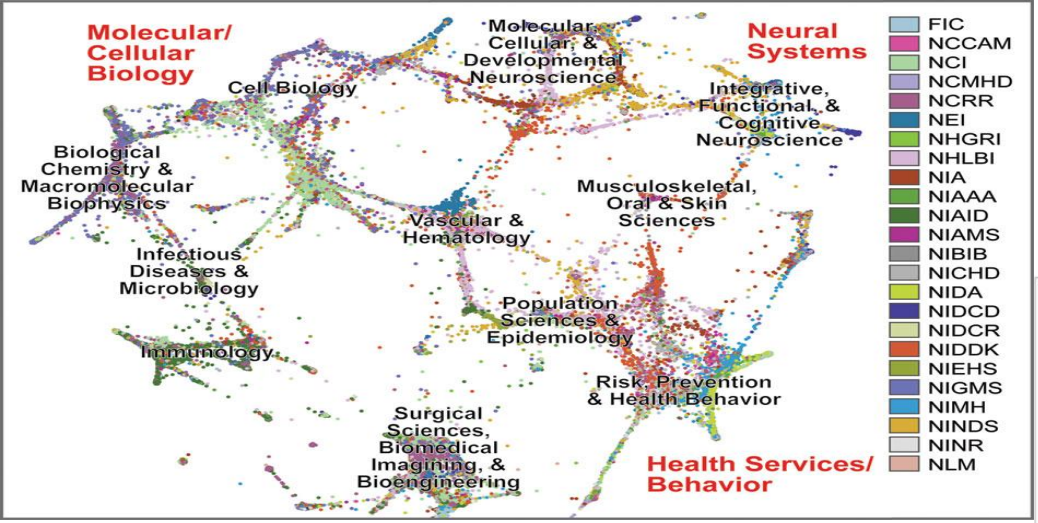
Document clustering is **“unsupervised” learning** in which there is no predefined category or “class,” but groups of documents that share similar topics are sought.

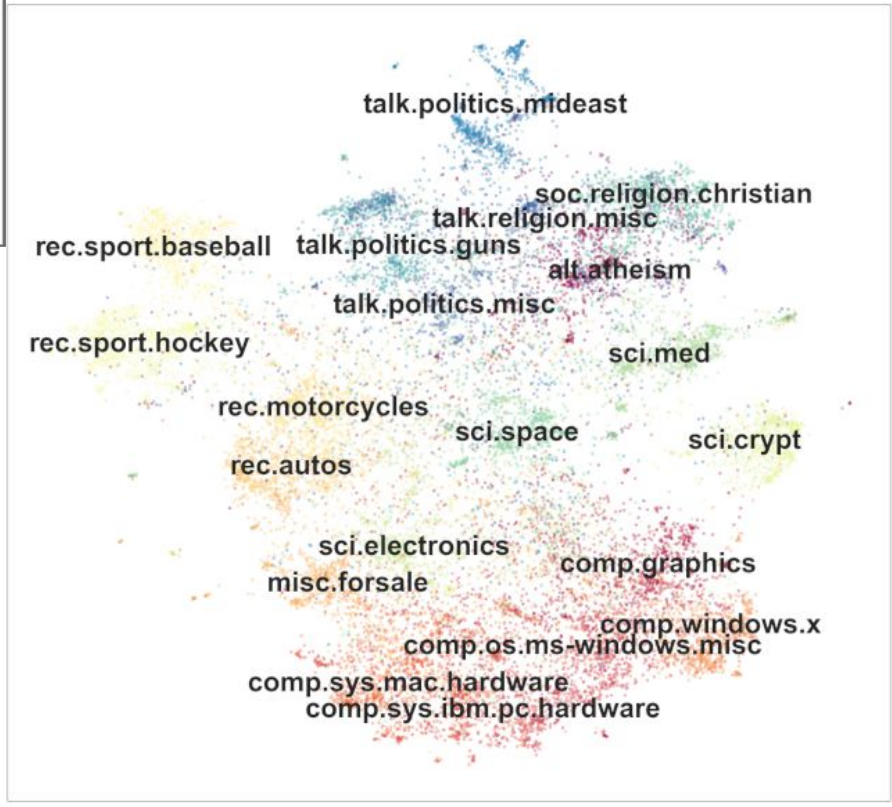
No Predefined Classes

Operates without explicit target labels, discovering natural groupings directly from raw document text.

Topic Similarity

Gathers documents into clusters based on shared thematic attributes, concepts, and content patterns.


Molecular/
Cellular
Biology
Molecular,
Cellular, &
Developmental
Neuroscience
Neural
Systems
Cell Biology
Integrative,
Functional, &
Cognitive
Neuroscience
Biological
Chemistry &
Macromolecular
Biophysics
Musculoskeletal,
Oral & Skin
Sciences
Vascular &
Hematology
Infectious
Diseases &
Microbiology
Population
Sciences &
Epidemiology
Immunology
Risk, Prevention
& Health Behavior
Surgical
Sciences,
Biomedical
Imaging, &
Bioengineering
Health Services/
Behavior
FIC
NCCAM
NCI
NCMHD
NCRR
NEI
NHGRI
NHLBI
NIA
NIAAA
NIAID
NIAMS
NIBIB
NICHD
NIDA
NIDCD
NIDCR
NIDDK
NIEHS
NIGMS
NIMH
NINDS
NINR
NLM


talk.politics.mideast
soc.religion.christian
talk.religion.misc
rec.sport.baseball
talk.politics.guns
alt.atheism
talk.politics.misc
rec.sport.hockey
sci.med
rec.motorcycles
sci.space
sci.crypt
rec.autos
sci.electronics
comp.graphics
misc.forsale
comp.windows.x
comp.os.ms-windows.misc
comp.sys.mac.hardware
comp.sys.ibm.pc.hardware

Text Mining Tasks | Topic Modeling

CORE TEXT MINING TASK

Topic Modeling

Topic modeling is the identification and tracking of the main topics in a document collection.

Topic Identification

Discovering latent thematic structures and extracting recurring subjects across large text document collections.

Topic Tracking

Monitoring how key topics evolve, shift in prominence, or emerge over time within a temporal document stream.

Topics

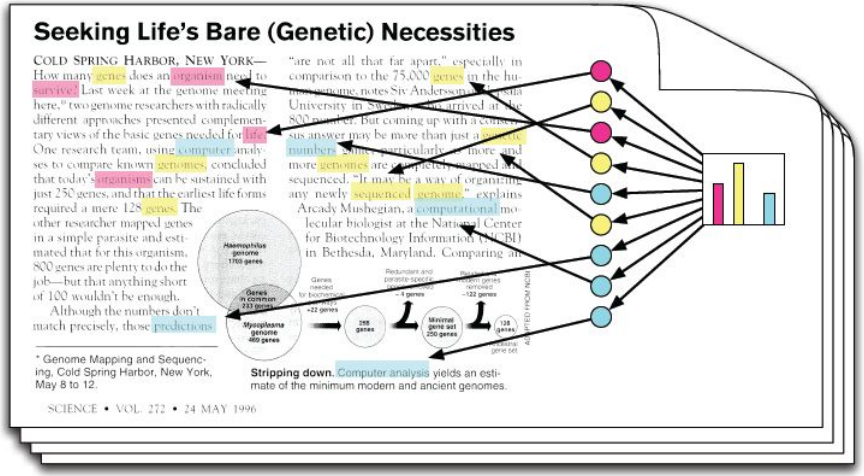
gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game, particularly as more and more genomes are completed and mapped or sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Topic proportions and assignments

Topic 1



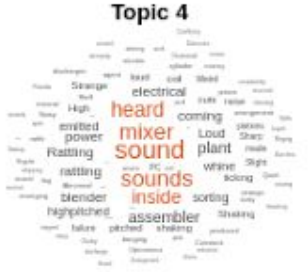
Topic 2



Topic 3



Topic 4



Text Mining Tasks | Text Summarization

CORE TEXT MINING TASK

Text Summarization

Text summarization strives to produce a condensed representation of its input, intended for human consumption.

Condensed Representation

Extracting essential information and reducing full source text length while preserving core semantic meaning.

Human Consumption

Formatting condensed output for rapid reading, high legibility, and effortless human comprehension.


strategy
processes
information
organization
technology
change
future
process
people
structure
align
planning
principles
delivery
effectively
objectives
alignment
models
blueprint
architecture
EA
capabilities
management
enable
discipline
strategic
systems
vision
goals
decisions
achieve
current
make
model
time
data
way
cost
development
knowledge
technologies
practice
realize
support
design
infrastructure
effective
software
changes
operating
approach
using
terms
defining
transformation
coherent
environment
best
ensure
technical
needs
supporting
informed
together
manage
set
efficiently
governance
used
body
Change
increase
ensuring
organizations
customer
investment
resources
execute
system
requirements
help
mind
mission
integration
state
view
value
meet
decision
enables
get
tools
Strategic
framework
reduce
stakeholders
execution
identity
desired
allows
aspects
operations
efficient
means
Planning
new
aligning
efficiency
right
structural
one
agility
opportunities
practices
Architecture
growth
strategies
organizational
Agility
standards
plan
improve
evolving
company
building
corporate
solution
create
making
enabling
assets
analysis
provides
components
functional
Efficiency
order
across
art
context
complexity
organization's
maintain
holistic
better
helps
services
organisation
transform
evolution

Text Mining
Methods and Approaches

Predictive & Exploratory Analysis of Texts

Predictive Analysis of Texts

Developing computer programs that automatically recognize or detect ***a particular concept*** within a span of text.

Exploratory Analysis of Texts

Developing computer programs that automatically discover ***interesting and useful patterns or trends*** in text collections.

How methods have changed over the years

Lexical features

Count words and phrases. Use TF–IDF and related **sparse representations.**

Learned representations

Use **contextual models** and sentence embeddings to capture some semantic similarity.

Generative models

Prompt or adapt language models to classify, extract, and summarize text.

Text Mining in Practice

Real-time Data Analytics (Relationship Modeling & Sentiment Analysis)

Text Mining in Practice – Understanding Consumer Intentions

What can social media posts tell us about a product?

Unsupervised consumer intention and sentiment mining from microblogging data as a business intelligence tool

Original Paper | Published: 12 May 2022

Volume 22, pages 6007–6036 (2022) Cite this article

Text Mining in Practice – Understanding Consumer Intentions

- **Our goal** was to support business decisions by analysing what people expressed about **a product** on Twitter.
- We examined **four intentions** — buying, not buying, comparing products, and seeking information.
- We also analyzed sentiment and whether posts expressed subjective or objective content.
  - Sentiment Mining
  - Opinion Mining
- We studied the **iPhone 5** to explore how these signals changed over time and around market events.

Text Mining in Practice – Our Approach

- Collect and filter posts. We extracted approximately **1.24 million English tweets mentioning the iPhone 5** from archived Twitter data and applied **filters** for **spam** and automated content (**Classifiers**).

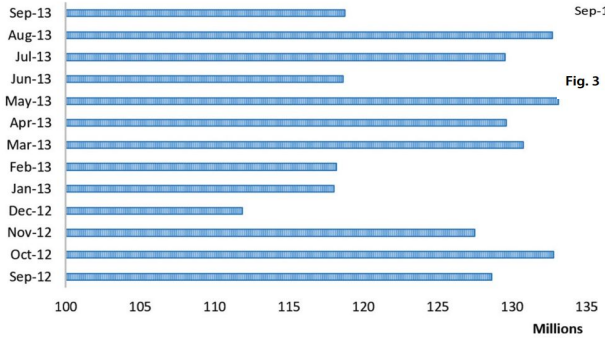


Fig. 2 Downloaded tweets per month

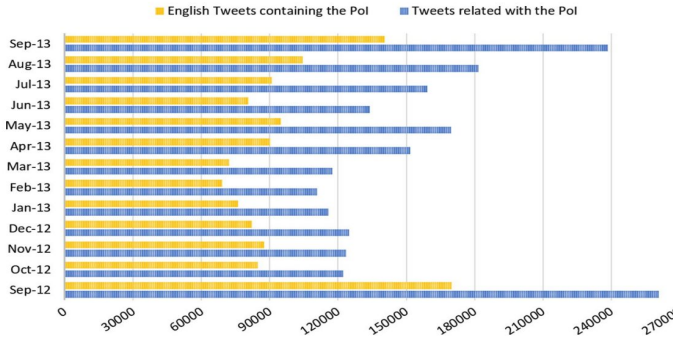


Fig. 3 Tweets related with the PoI and English Tweets containing the PoI

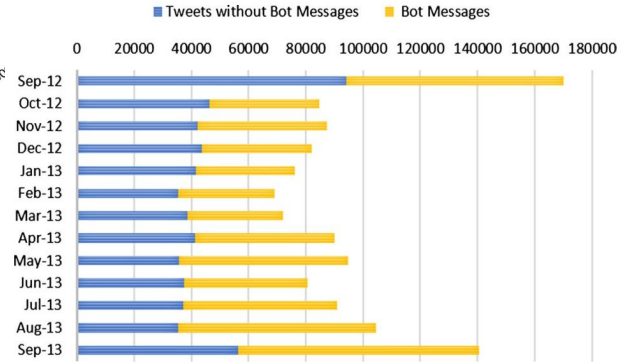


Fig. 4 Bot messages and final set of Tweets

Mining Consumer Intentions – Intention to (not) Buy iPhone 5

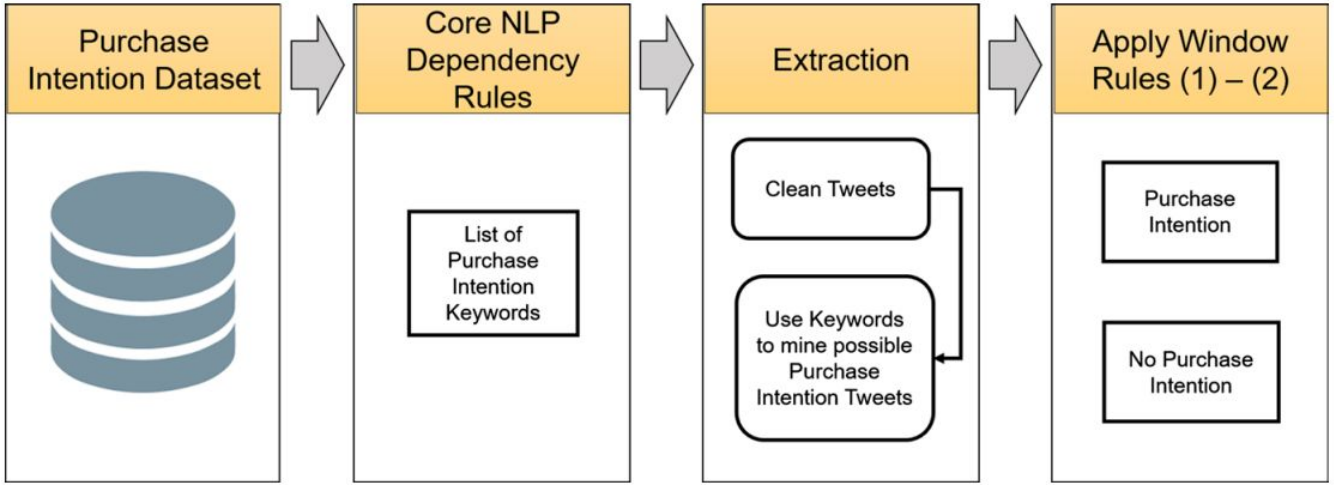


Fig. 5 Extraction methodology for purchase intention

Table 3 Purchase intention verbs and phrases

Verbs	Looking, get, buy, want, need, purchase, use, replace, choose, take, find, like, thinking
Phrases	Would like, should buy, would choose, could get, could find, would use, would have

Mining Consumer Intentions – Intention to (not) Buy iPhone 5

Table 4 Examples of tweets with purchase intention

Tweet
I swear i want that new iphone
RT @USER: I need the iphone 5 in my life.
Finna get an iPhone this week.. yuppp
I m be getting da iPhone 5s beofe ya'lall

Mining Consumer Intentions – Intention to Compare iPhone 5

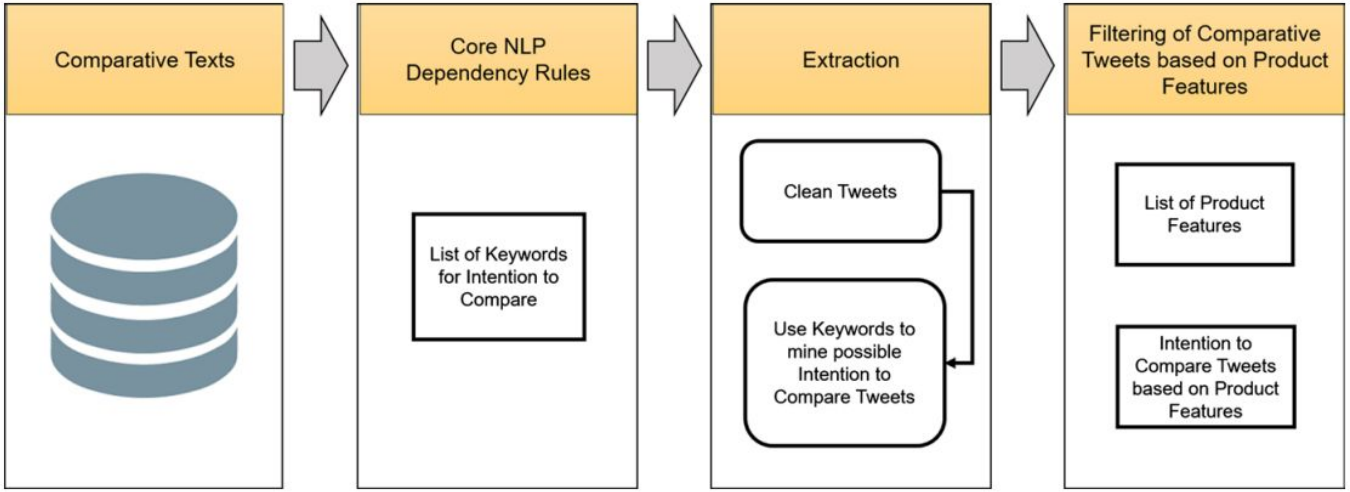


Fig. 6 Extraction methodology for compare intention

Table 5 Keywords to extract comparative user intentions

Keywords	Prefer, fewer, as easy as, comparable, as good as, one of the, higher, better, exceeds, compared to

Mining Consumer Intentions – Intention to Compare iPhone 5

Table 6 Examples of tweets with comparative intention

Tweet
I know I've tweeted this before, but I really need the iPhone 5 to come out.
My current phone is sad.
RT @USER: Need an iphone asap, cannot deal with this blackberry any longer.
@USER your commercials are so bad and your attempts to slander the iPhone make you look pathetic. Just accept 2nd place... At best.
I dont know why twitter on iPhone is much cooler than on android :p

Table 7 List of iPhones 5 comparative products

Comparative products	Blackberry, motorola, sony experia s, htc, nokia lumia, lg nexus, samsung

Mining Consumer Intentions – Intention to Seek Information iPhone 5

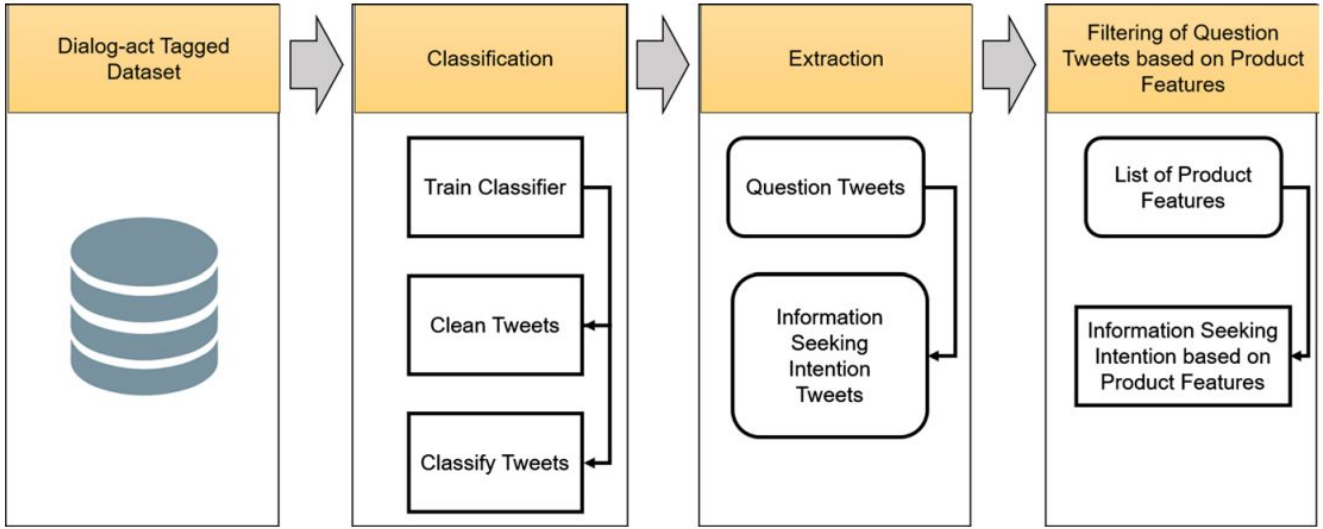


Fig. 7 Extraction methodology for seeking information intention

Table 8 List of iPhone 5 features

iPhone's 5 features	Price, camera, applications, iOS battery, screen, case, warranty

Mining Consumer Intentions – Intention to Seek Information iPhone 5

Table 9 Examples of Tweets with seeking information intention

Tweet
Looking for new iPhone cases.. Any suggestions?
Is iPhone 5 battery "long lasting" enough?
Why does the front camera of an iPhone have to be low quality?
So what's new with the iPhone 5?

Mining Consumer Intentions – Sentiment & Opinion Classification

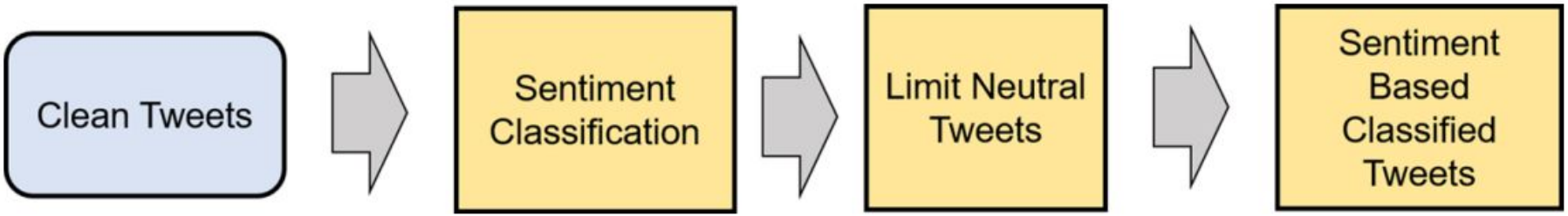


Fig. 8 Sentiment polarization detection

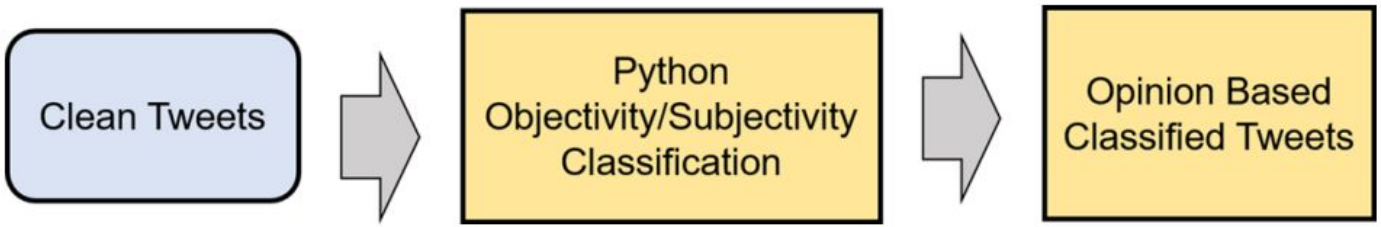


Fig. 9 Opinion classification

Mining Consumer Intentions – Sentiment & Opinion Classification

Table 10 Examples of Tweets with sentiment and opinion

Sentiment	Opinion	Tweet
Negative	Subjective	The only thing I don't like about my iPhone is how It always deletes my contacts.
Positive	Subjective	@USER I can't take this phone any more . And I like all the cases and the messaging app. Hopefully ill get the iPhone 5
Neutral	Objective	does anyone know if there's an untethered jailbreak for iOS 6.0.1 for the iphone 4???
Negative	Objective	it annoys me how i can't see iphone emoticons on google chrome

Results – **Predictive** Business Insights

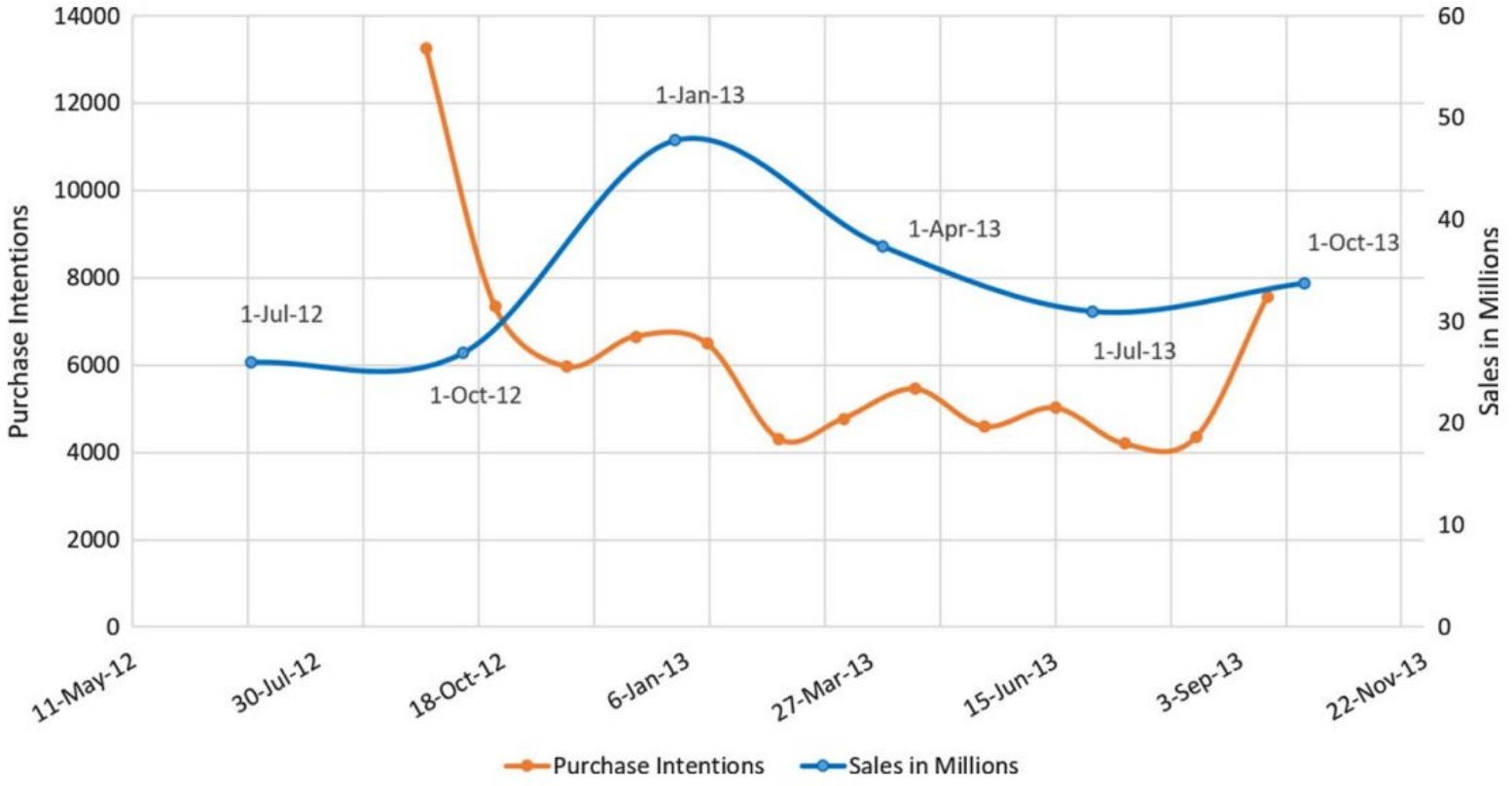


Fig. 11 Relationship between intentions of purchase and number of sales

Results – Business Insights

- Mining Intentions, given an business triggered international event.

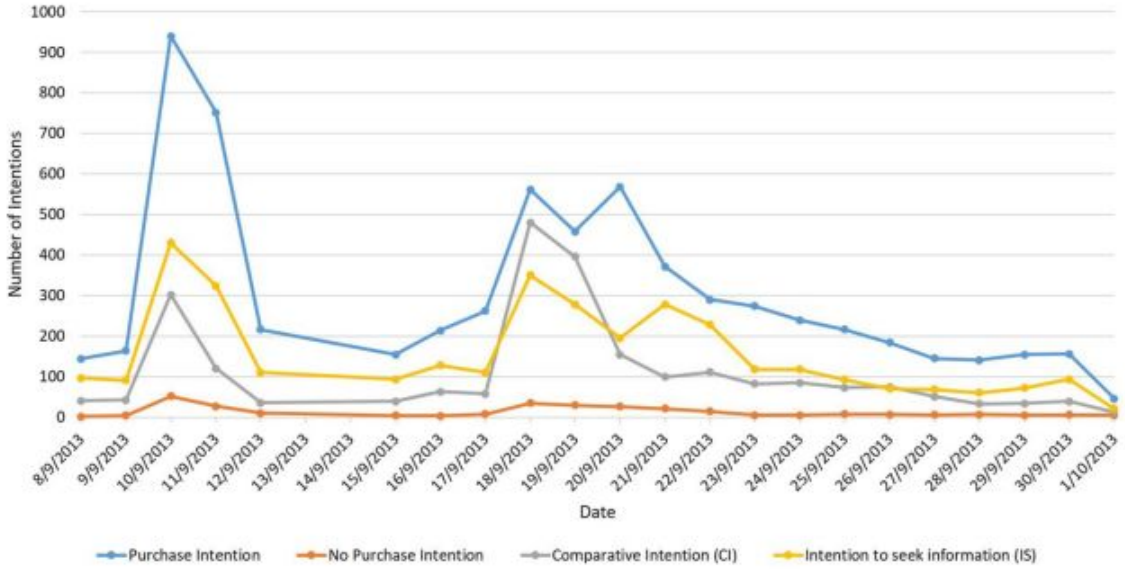


Fig. 19 Mining of intentions during the announcement of Apple iPhone 5c, 5s models (10-September-2013)

Table 11 Case study's example Tweets

Intention category	Tweet
Comparative intention	Samsung S3 + iPhone are more or less the same now! But the S3 got alot more things i prefer! E.g. BlueTooth!! Battery Life + easier 2 use
Purchase intention	I already got a nexus. i want an rbz. but i wanna buy the iphone 5 or a samsun
Information seeking intention	Which one should I choose iphone or S3?

Results – **Exploratory** Business Insights

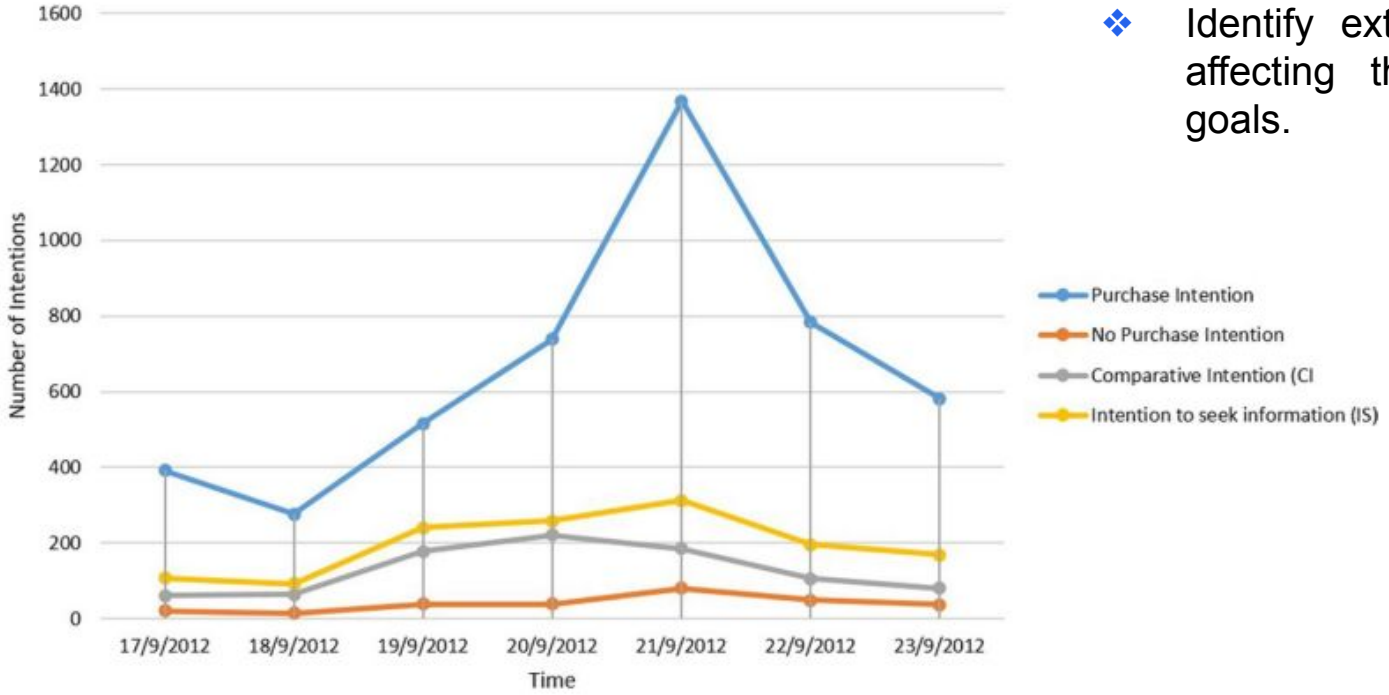


Fig. 18 Intent extraction in Samsung's aggressive campaign

- Identify external actions that are affecting the business and our goals.

Text Mining
Examples & Challenges

Six reviews, several questions

R1

The pasta was excellent, but we waited 40 minutes.

R2

Friendly staff. My pizza arrived cold.

R3

Our meal arrived quickly and tasted great.

R4

Not bad food, but the bill included an extra drink.

R5

Wonderful. Another hour waiting for a table.

R6

The staff corrected the bill immediately.

What should the restaurant investigate first?

A task is defined by its output, often we need to mix.

Classification

Assign a review to a known category.

Extraction

Identify the dish, price, or reported waiting time.

Clustering

Group reviews with similar content.

Summarization

Produce a shorter account of the evidence.

“Understand our customers” is a goal. It still needs an explicit task.

Exercise 1 - A small annotation task

Which reviews explicitly describe a waiting problem?

1. Work in pairs. Read R1–R6 on slide 2.
2. Mark each review Yes, No, or Uncertain.
3. Underline the words that support each decision.
4. Compare one disagreement and refine your labeling rule.

Six minutes, including discussion.

Language makes the task difficult

Negation	“The food was not good.”
Sarcasm	“Wonderful. Another hour waiting for a table.”
Different wording	“Slow service” and “We waited forever.”
Missing context	“It was better this time.”

The same words can carry different meanings in different contexts.

The corpus defines what we can “mine” from data

Collection choices

Which restaurants and platforms?

Which dates and languages?

One review or one sentence per item?

Potential distortions

Duplicate reviews inflate counts.

Missing reviews change the sample.

Frequent reviewers can dominate.

A pattern in the collected text may not represent all customers.

Challenges in Text Mining

Unstructured Text Format

Documents in raw textual form are not readily accessible or structured for direct processing and computation by machines.

Massive Scale & Volume

Efficiently managing and analyzing massive ***collections of documents*** or continuous ***streams of text***.

Lack of Organization

Textual data is rarely organized **neatly into tables,** existing predominantly as semi-structured or completely unstructured content.

Multilevel Ambiguities

Natural language contains inherent ambiguity across multiple linguistic levels: **Lexical, Syntactic, Semantic, and Pragmatic**.

A Text Mining Workflow

01

Define the question

Choose the output and how it will be used.

02

Build the dataset

Collect text and specify labels or evidence.

03

Represent the text

Create counts, features, or embeddings.

04

Apply a method

Fit a model or use a specified procedure.

05

Evaluate & inspect

Measure errors and revise the design.

The process is iterative. Error analysis can send us back to the data.

Contacts

Dr. Georgios Peikos

georgios.peikos@unimib.it

Academic Year 2026/27

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Ikr3 Lab — information and knowledge representation, retrieval, and reasoning LAB