INTRODUCTION TO TEXT MINING

Marcus Birkenkrahe

January 7, 2023

README

You will learn:

□ The basic definition of practical text mining

□ Why text mining is important to the modern enterprise

□ Examples of text mining used in enterprise

□ The challenges facing text mining

□ Example workflow for processing natural language

□ A simple text mining example

□ When text mining is appropriate

Source: Kwartler, 2019, chapter 1

What is "text mining"?

Figure 1: Four monks by Claudio Rinaldi (1852-1909), Dorotheum, Munich

> "Text mining is the process of distilling actionable insights from text." Kwartler, 2019

□ What does "distilling actionable insights" mean?[1]

> "Text [data] mining is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights." IBM, 2023

□ What does "structured" and "unstructured data" mean? What are "meaningful patterns"?[2]

[1] Distillation is a process of extracting an essence (a wanted substance) and getting rid of unwanted substances. Actionable insights are insights that one can use to make decisions (action in business is usually accompanied by decision-making).

[2] (Source) Structured data are data in tabular format with specific data types for digital processing. Unstructured data do not have a specific data format.

- □ What about the text on this page: structured or not?[3]

What is text mining in practice (= business)?

- Identify useful social media posts for customer services
- Measure campaign success for marketing purposes
- Match job descriptions to resumes for human resources

Why should you care about text mining?

> " To be truly customer-centric in a hyper-competitive environment, an organization should be listening to their constituents whenever possible. Yet the amount of textual information from these interactions can be immense, so text mining offers a way to extract insights quickly." Kwartler, 2019

Concrete examples:

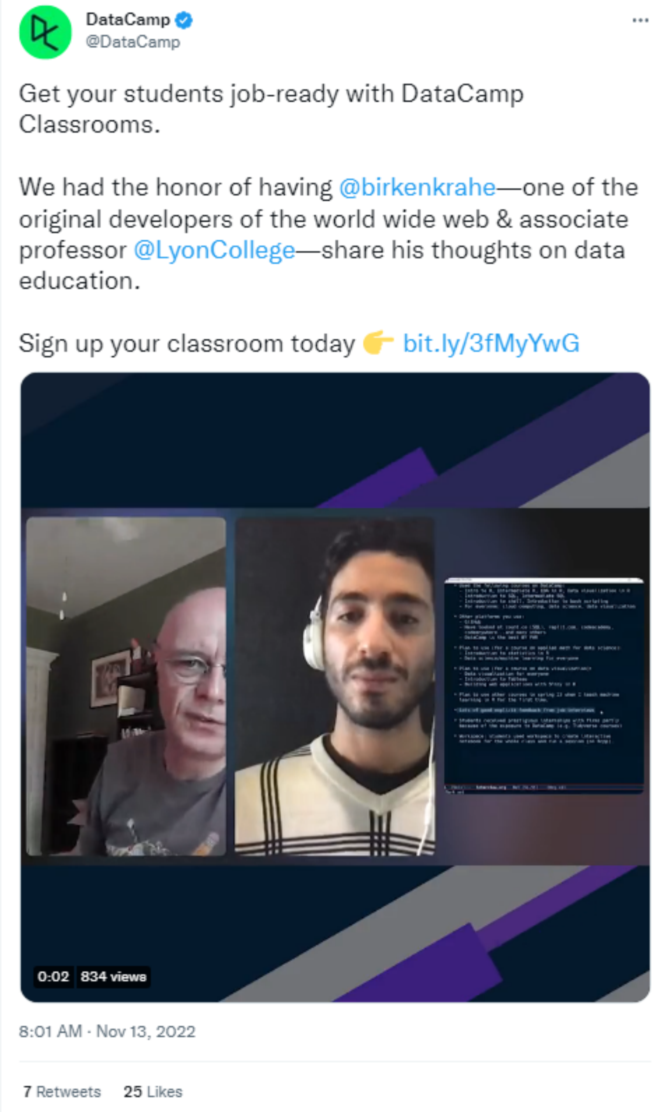
- Relevance of social media for public opinion (e.g. Twitter)
- Growth of online content from an organization (e.g. blogs)
- Digitization of paper records (e.g. healthcare)
- Automatic translation of natural language (e.g. Google Translate)
- Augmentation of human work through chatbots (e.g. ChatGPT)

Vox populi - the "wisdom of crowds"

> "Under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them." Surowiecki, 2005

The "right circumstances": no assessment bias =

[3] The Org-mode file is semi-structured! Semi-structured data carry meta information in the form of markup - e.g. HTML, XML, JSON, or Org-mode: the header information at the top of the file structures the data, as does the Org-mode format itself, which comes with a markup language.



DataCamp
@DataCamp

Get your students job-ready with DataCamp Classrooms.

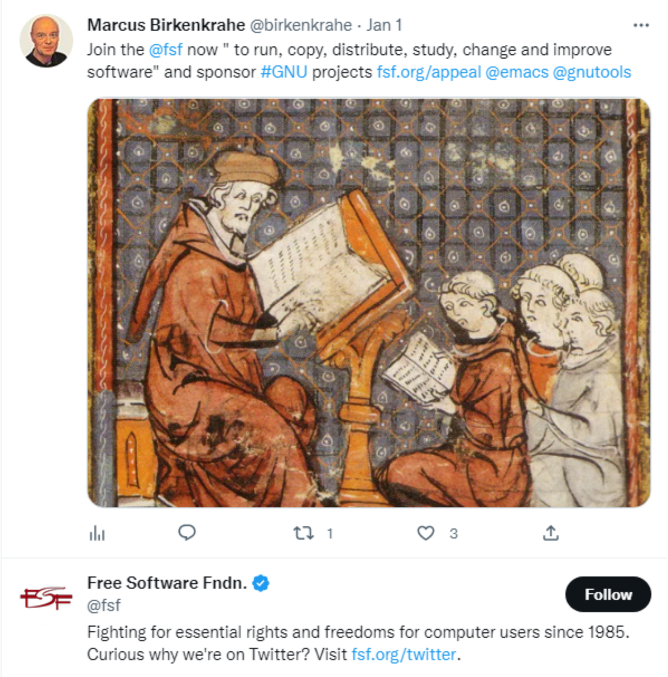
We had the honor of having @birkenkrahe—one of the original developers of the world wide web & associate professor @LyonCollege—share his thoughts on data education.

Sign up your classroom today 👉 bit.ly/3fMyYwG

8:01 AM · Nov 13, 2022

7 Retweets 25 Likes

Figure 2: twitter.com/DataCamp featuring interview with me



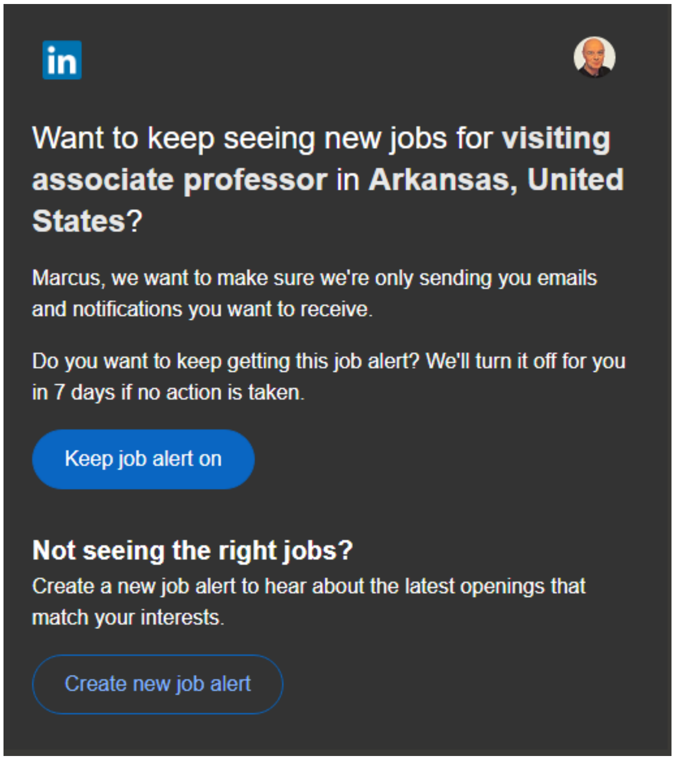
Marcus Birkenkrahe @birkenkrahe · Jan 1

Join the @fsf now " to run, copy, distribute, study, change and improve software" and sponsor #GNU projects fsf.org/appeal @emacs @gnutools

Free Software Fndn.
@fsf

Fighting for essential rights and freedoms for computer users since 1985. Curious why we're on Twitter? Visit fsf.org/twitter.

Figure 3: twitter.com/birkenkrahe campaigning for fsf.org



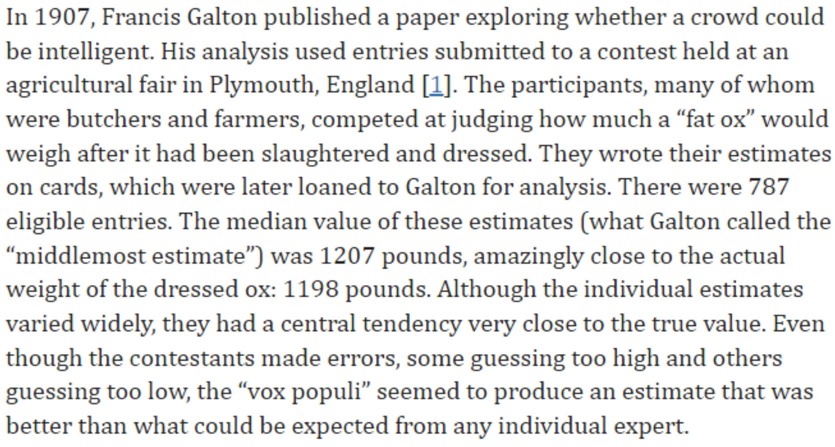
Want to keep seeing new jobs for **visiting associate professor** in **Arkansas, United States**?

Marcus, we want to make sure we're only sending you emails and notifications you want to receive.

Do you want to keep getting this job alert? We'll turn it off for you in 7 days if no action is taken.

Not seeing the right jobs?
Create a new job alert to hear about the latest openings that match your interests.

Figure 4: automatic job alert by LinkedIn.com



In 1907, Francis Galton published a paper exploring whether a crowd could be intelligent. His analysis used entries submitted to a contest held at an agricultural fair in Plymouth, England [1]. The participants, many of whom were butchers and farmers, competed at judging how much a "fat ox" would weigh after it had been slaughtered and dressed. They wrote their estimates on cards, which were later loaned to Galton for analysis. There were 787 eligible entries. The median value of these estimates (what Galton called the "middlemost estimate") was 1207 pounds, amazingly close to the actual weight of the dressed ox: 1198 pounds. Although the individual estimates varied widely, they had a central tendency very close to the true value. Even though the contestants made errors, some guessing too high and others guessing too low, the "vox populi" seemed to produce an estimate that was better than what could be expected from any individual expert.

Figure 5: Source: Patten, 2015

1. Assessors need to exercise *independent* judgements

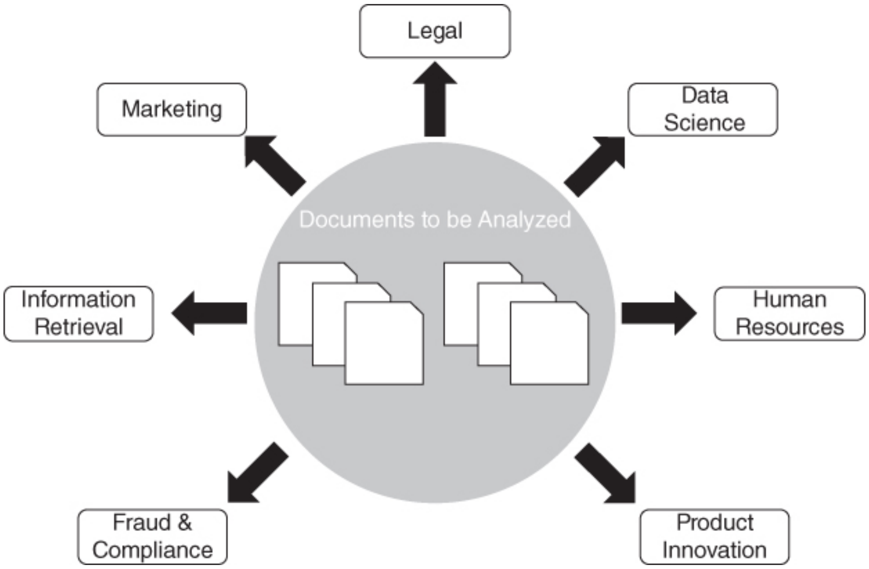
2. Assessors need to possess *diverse* information understanding

3. Assessors need to rely on *decentralized*, *local* knowledge.

4. There has to be a way to *aggregate* or tabulate the results.

5. □ How about Amazon.com reviews - do they meet these conditions?[4]

Beneficiaries and benefits of text mining



- Benefits include:
  1. Trust among stakeholders because little to no *sampling* is needed to extract information (all available text sources can be used).
  2. The methodologies can be applied quickly (text processes fast).

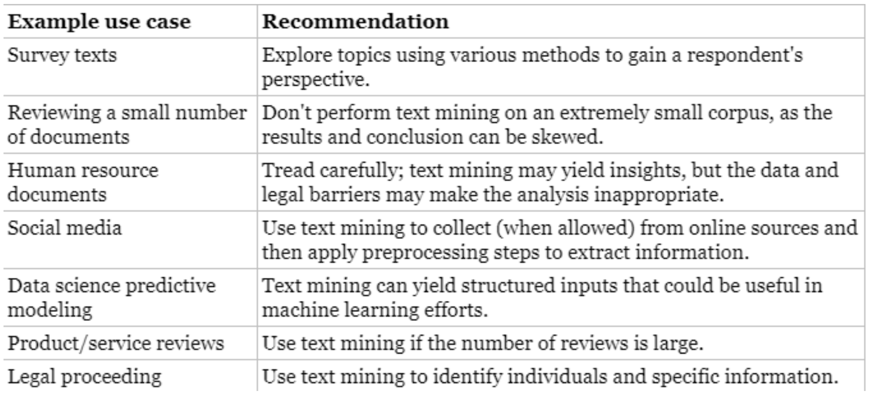
[4] (1) reviews may not be independent since reviewers have access to old reviews, which may influence them (it's harder to have a different opinion from everyone else). (2) Diversity is hard to measure but in the case of Amazon.com, a national audience can be seen as highly diverse (there are nearly 150 mio subscribers of Amazon Prime in the US alone). (3) Local here means "not only at a distance" - only "verified purchase" reviews fulfil this condition in principle. (4) Tabulation of the reviews relies on text mining, and hence - unlike in the case of Galton - not on recording simple numbers. Stochastic procedures (probability distributions) are involved.

3. Using R allows for *auditable* and *repeatable* methods.
4. Text mining identifies novel *insights* or reinforces existing perceptions based on all relevant information.

- The "opinion" of ChatGPT looks comprehensive as always - does this chatbot represent "vox populi"? Are all criteria fulfilled?[5]

□ Whom would you trust more - the expert author or the chatbot?[6]

When to use and when not to use text mining



Example use case	**Recommendation**
Survey texts	Explore topics using various methods to gain a respondent's perspective.
Reviewing a small number of documents	Don't perform text mining on an extremely small corpus, as the results and conclusion can be skewed.
Human resource documents	Tread carefully; text mining may yield insights, but the data and legal barriers may make the analysis inappropriate.
Social media	Use text mining to collect (when allowed) from online sources and then apply preprocessing steps to extract information.
Data science predictive modeling	Text mining can yield structured inputs that could be useful in machine learning efforts.
Product/service reviews	Use text mining if the number of reviews is large.
Legal proceeding	Use text mining to identify individuals and specific information.

- "Use case": an application scenario used for illustration

- Lists should always be ordered (explicitly or implicitly)

□ How could one order the list of example use cases?

[5] ChatGPT is source from a very large number of textual documents but it is impossible to ascertain any of the criteria when identfying the chatbot as the "assessor".

[6] For me personally, knowledge about a source increases trust in believing that source while lack of knowledge decreases the trust. In the case of ChatGPT, I asked the bot about its sources but its answer was redundant and not overly satisfying (see for yourself).

Language is not like other data

□ What is special about language data?[7]

- "The true origin of language may never be known." (ChatGPT)
- Text mining reduces the information available in language

Avoid word clouds - beware of the cliché

- Use them in conjunction with other methods to confirm the correctness of a conclusion

□ What do you think why word clouds are still so attractive?

Basic text mining workflow

1. Define the problem and specific goals (e.g. how best to market)

[7] Language is used for communication; it is thought to be divine or at least strongly linked to the divine ("In the beginning was the Word, and the Word was with God, and the Word was God." John 1:1); it may be that only humans have language; it is learnt.

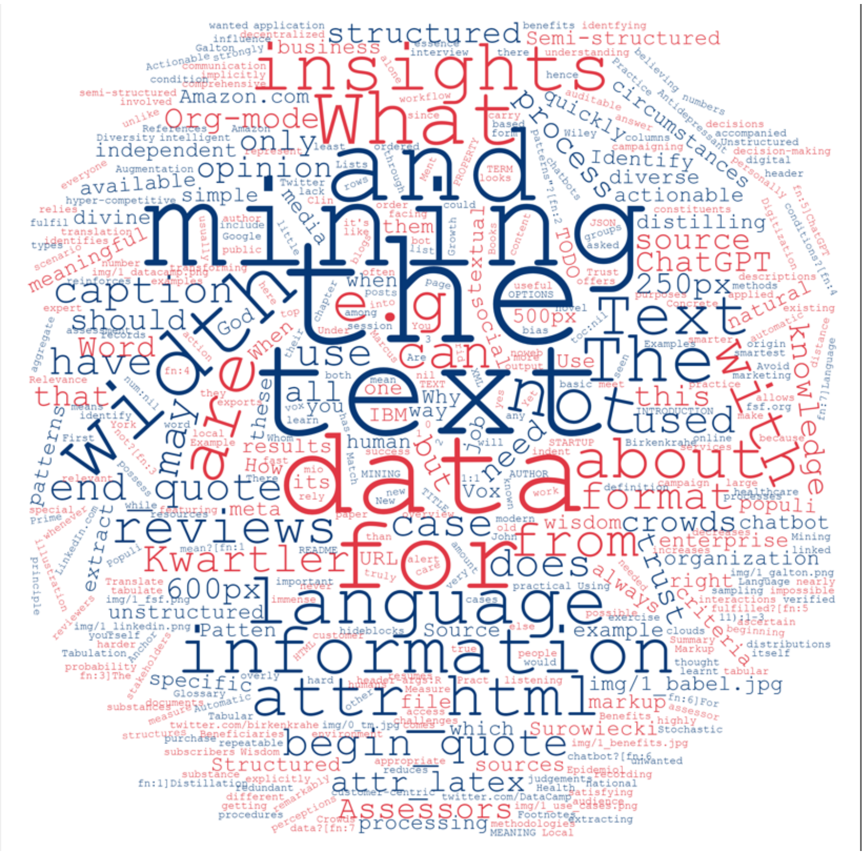


Figure 6: Wordcloud on the words of this lecture - wordclouds.com

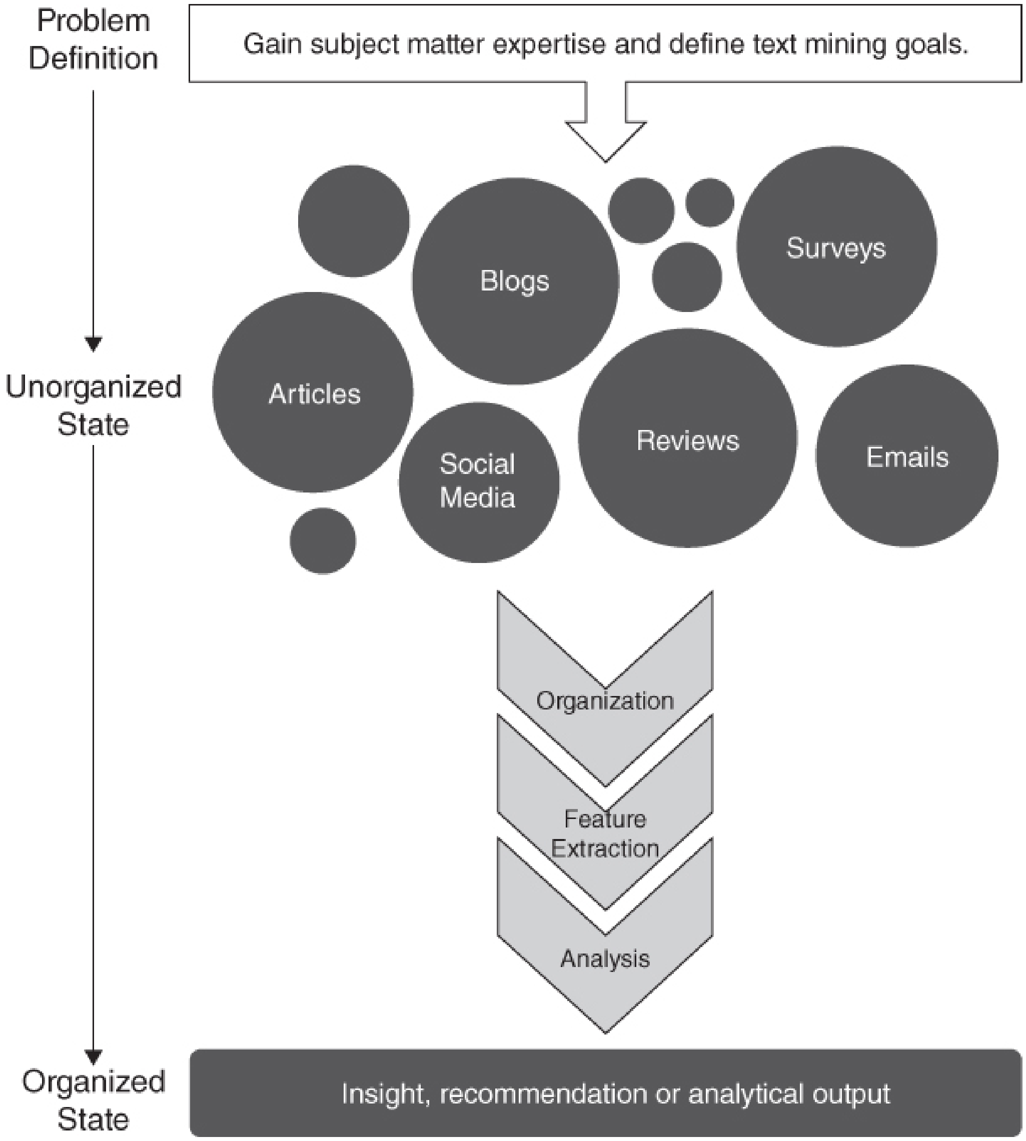


Figure 7: Source Kwartler (2019)

2. Identify the text that needs to be collected (e.g. Twitter API)
3. Organize the text (e.g. into a corpus for "bag of words")
4. Extract features for analysis (e.g. make text lower case)
5. Apply techniques to the prepared text (e.g. keyword search)
6. Reach an insight or recommendation (e.g. marketing focus)

Which tools are needed?

- Sufficient RAM for R (all processing is done in memory)
- Installation of R and an IDE like RStudio or Emacs + ESS + Org-mode
- Set of R packages and example data
- Any operating system (Linux is to be preferred)

Simple example: mining customer reviews

- You're a Nike employee who wants to know how consumers are viewing the Nike Men's Roshe Run Shoes. Follow these steps:

  1. Goal definition: Using online reviews, identify overall positive or negative views. For negative views, identify cause to be shared with the product manager.
  2. Data collection: For a mass market product, use retail website like Amazon for hundreds of timestamped reviews (to ensure currency).
  3. Text organisation: Web scrape all reviews into a CSV file with one review per row, timestamp and star rating to later subset corpus by these features.

4. Feature extraction: clean reviews to analyze text features, e.g. removing common words with little benefit ("shoe", "nike", "running" etc.). Check for spelling and make all text lowercase.
5. Text analysis: scan for specific group of keywords depending on product issues ("fit", "rip", "tear", "narrow", "wide", "sole"). Sum group counts to order problematic features.
6. Insight generation: present findings to product manager that the top consumer issue is "narrow" and "fit" to aid product design, marketing or improvement decisions.

Real world example: competitive intelligence

- Text mining can help to understand the basics of a competitor's text based marketing (for further analysis, contrast or imitation)
- When creating Amazon.com's social customer service team, they were "obsessed with how others were doing it".
- They read and reviewed other companies customer replies and learnt from their missteps.[8]
- In 2012, social media based customer service was considered to be highly risky, involving legal counsel, branding, and leadership.
- In 2012, Wal-Mart, Dell and Delta Airlines were considered best in class social customer service companies.
- Each brand owner (Amazon Prime, Amazon Kindle etc.) had cultivated their own style of communicating via social media (like dialects).
- Every communication channel was supposed to execute flawlessly and be 100% customer-centric.
- Goal: develop social media cautiously to maintain current quality set by multiple stakeholders.
- Initial channels: two help forums, retail and Kindle Facebook pages and Twitter.

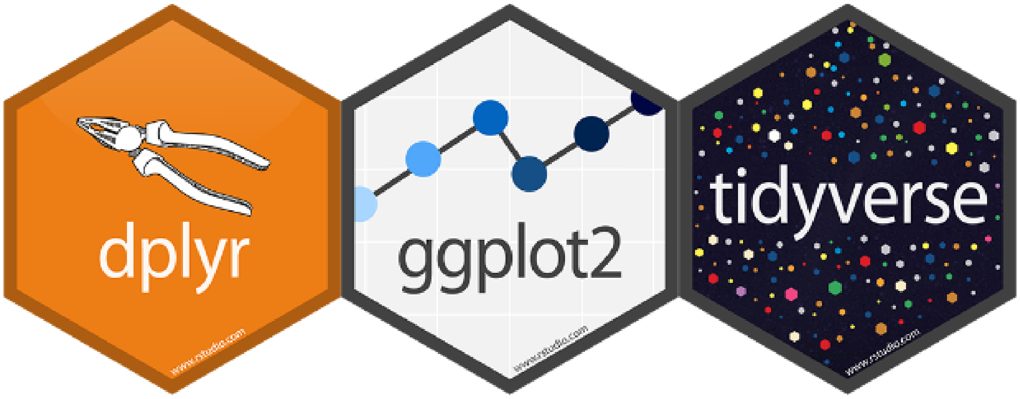
[8] This reminds me of my own experience with CISCO customer services when working at Shell and visiting CISCO to (openly) learn from their knowledge sharing experiences.

- Text mining was a tool to analyze competitors' use of social media for customer services: grasp length of a reply (e.g. Twitter limit), language used, typical customer agent workload, and if posting similar links repeatedly made sense, what types of help links to post (forms, resource links?), how many people should be doing this, etc.
- Text mining focused on three questions for about one year:
  1. What is the average length of a social customer service reply?
  2. What links were referenced most often?
  3. How many social replies is reasonable for a customer service agent to handle?
- By 2017, Amazon was a leading force in this space (WBR, 2023)

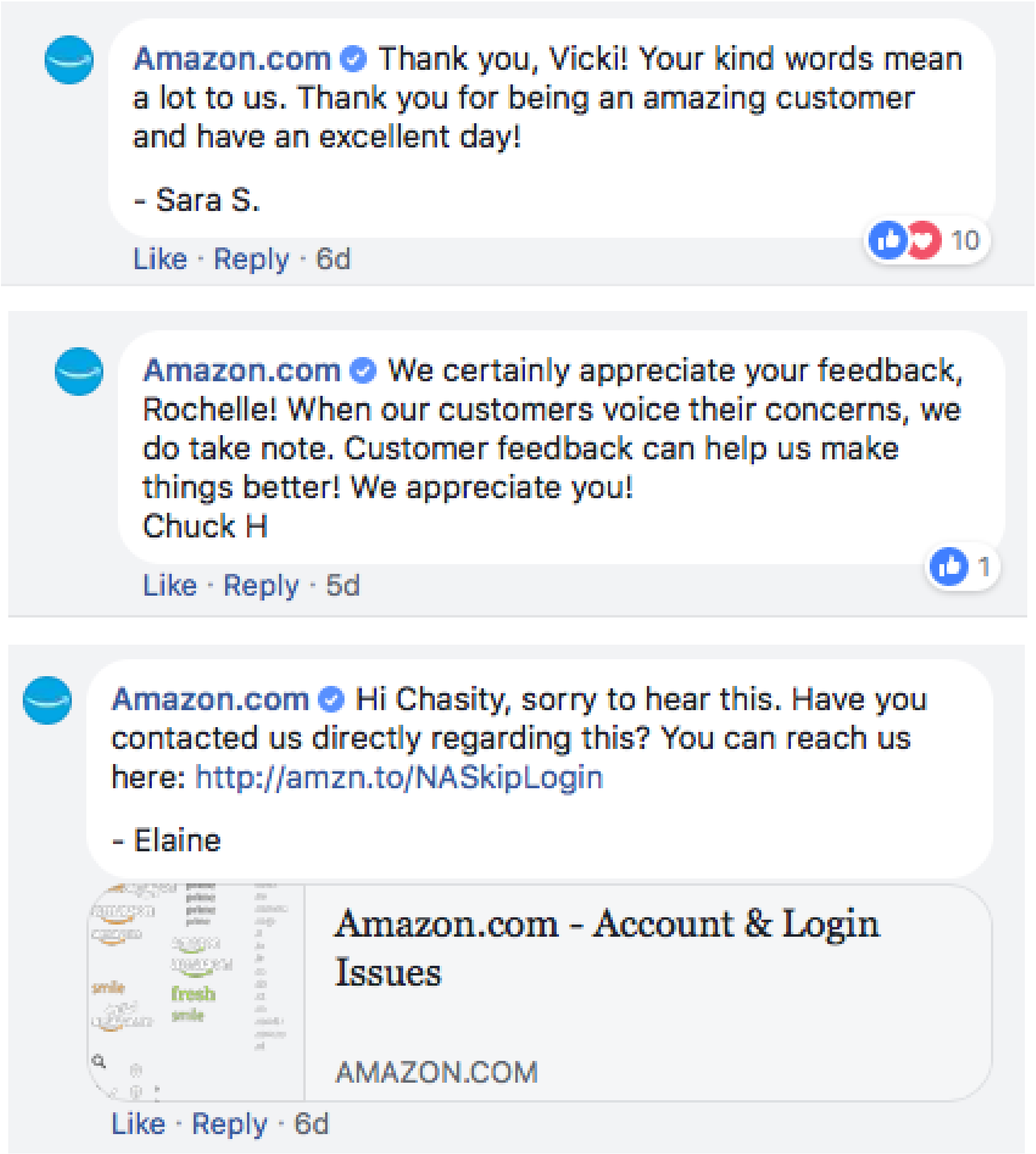
Final definition for "text mining"

> "Text mining represents the ability to take large amounts of unstructured language and quickly extract useful and novel insights that can affect stakeholder decision-making."

Next



Introduction to R and the tidyverse to make you fit for the first DataCamp challenge.



Amazon.com ✓ Thank you, Vicki! Your kind words mean a lot to us. Thank you for being an amazing customer and have an excellent day!

- Sara S.

Like · Reply · 6d

10

Amazon.com ✓ We certainly appreciate your feedback, Rochelle! When our customers voice their concerns, we do take note. Customer feedback can help us make things better! We appreciate you!
Chuck H

Like · Reply · 5d

1

Amazon.com ✓ Hi Chasity, sorry to hear this. Have you contacted us directly regarding this? You can reach us here: http://amzn.to/NASkipLogin

- Elaine

Amazon.com - Account & Login Issues

AMAZON.COM

Like · Reply · 6d

Figure 8: Amazon social media customer service examples (Facebook)

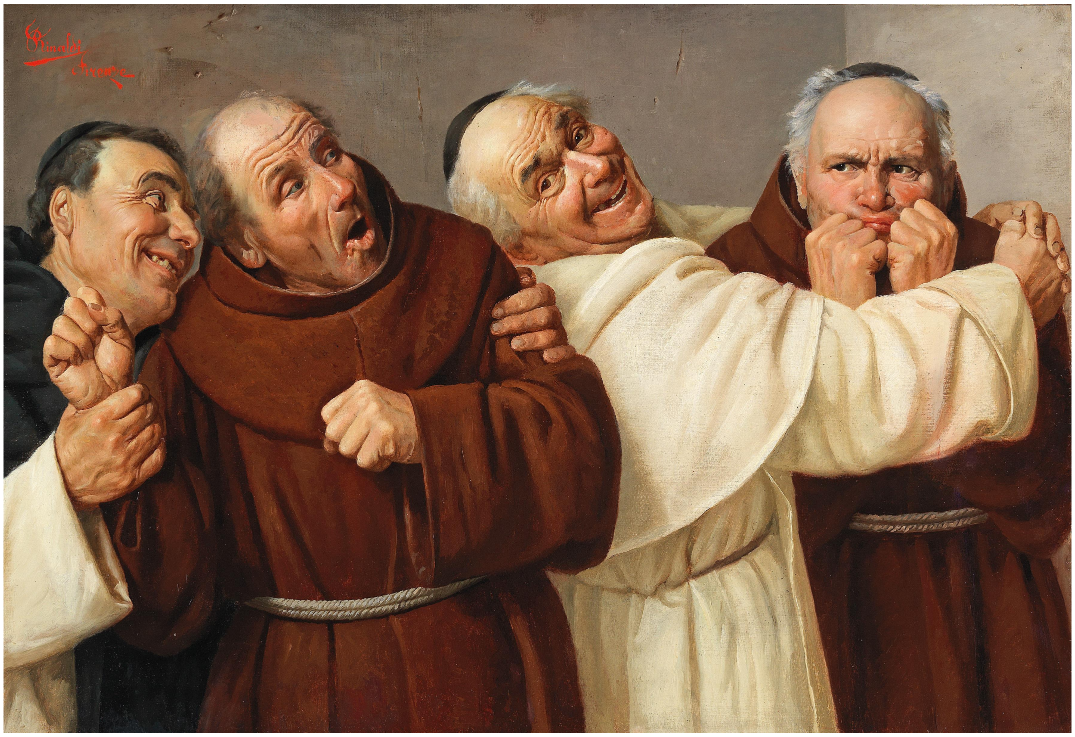


Figure 9: Four monks by Claudio Rinaldi (1852-1909), Dorotheum, Munich

TM Glossary

TERM	MEANING
Text mining	Identify useful patterns in text
Structured data	Tabular data (rows and columns)
Semi-structured data	Markup with meta data
Wisdom of crowds	Intelligence exhibited by groups
Use case	Illustrative application scenario
Feature extraction	Preprocess text for analysis
Corpus	Body of text to be analyzed
Stakeholder	Someone who cares
Competitive intelligence	Information about one's competitors

References

- IBM (2023). What is text mining? URL: ibm.com/topics/text-mining.

- Kwartler, T (2019). Text Mining in Practice with R. Wiley.

- Patten, S B (2015). The Wisdom of Crowds (Vox Populi) and Antidepressant Use. Clin Pract Epidemiol Ment Health (11):1-3. URL: doi.org/10.2174%2F1745017901510011001

- Surowiecki J (ed) (2005). The wisdom of crowds. New York First Anchor Books. crowds.