

VOLUME 1 | ISSUE 3 | JULY 2026

# THE DATA DIALOGUE

Your Weekly Deep Dive into  
Insights, Analysis & Trends

# The Data Dialogue.

*Simplified. Not Simplistic*

July. 29<sup>th</sup>. 2026 | Vol.1 | Issue 3 | Intro: Text Mining

## *In-Focus*

### *Grab the Reins: The Rise of Text Mining*

In a famous 2011 study published in *Science*, Jean-Baptiste Michel, Erez Lieberman Aiden, and the Google Books Team used text mining to track human history. By scanning 5.2 million digitized books, they measured the rise and fall of celebrity fame over time. They concluded that modern celebrities become famous much faster than in the past, but the public also forgets them much



(So, cheer up when your next video or post does not become viral, you are on the ancestor's right track)

Ultimately, this study proved that text mining is a brilliant way to gather data. Instead of forcing humans to read thousands of pages, computer programs instantly turn messy text—like old archives, emails, or medical notes—into neat spreadsheet data. Unrelated and unstructured data suddenly became relatable, structured, and can have meaning. Beyond numbers and narratives, text mining serves as a powerful bridge between two traditionally separate worlds: qualitative and quantitative research. By automating the analysis of thousands of open-ended text responses, researchers no longer have to choose between the deep context of personal stories and the statistical power of big data. Text mining champions this mixed-methods approach by allowing you to extract hard numbers—like word frequencies or sentiment scores—from highly descriptive narratives. It proves that you can run rigorous statistical tests on data without ever losing the human voice behind the words.

By finding patterns that standard questionnaires miss, text mining lets us read between the lines of history at lightning speed, turning millions of old words into useful clues about human behavior. If you are a qualitative researcher, next time when your quantitative fellows jump on their high horse, jump higher and grab the reins because text mining is leading the way now.

# The Data Dialogue.

*Simplified. Not Simplistic*

July. 29<sup>th</sup>. 2026 | Vol.1 | Issue 3 | Intro: Text Mining

## Tools

### *Two Ways of Text Mining*

#### 1. Code-Dependent Packages in R & Python

##### 1.1. For R

###### 1.1.1. Quanteda

##### 1.2. For Python:

###### 1.2.1. Natural Language Toolkit

###### 1.2.2. spaCy

###### 1.2.3. Gensim

#### 2. Visual Platform with No Code Requirements

##### 2.1. Orange Data Mining

##### 2.2. KNIME

##### 2.3. RapidMiner

## *Pitfall Patrol*



using the command `attach ()` in R.



**Instead**, do not use `attach ()` and depend on calling the variables using the dollar sign and the data frame name

**In case you have already used it**, you can undo the attach by using the command `detach ()`

**Why?** Beginners think that using the `attach ()` command once will save them time and spare them using the name of the dataset/ data frame each time. However, practice shows that this command creates more problems than expected: masking variables when you work with more than data frame, creating error messages, and sometimes works with read-only version without real execution on the data. To understand more, read the following example:

# The Data Dialogue.

*Simplified. Not Simplistic*

July. 29<sup>th</sup>. 2026 | Vol.1 | Issue 3 | Intro: Text Mining

Imagine that R environment is an office with two teams (data frames): the Marketing team (Marketing) which has two members (Variables): Year and Facebook Followers. The Second is Sales team (Sales) which has two members (Variables): Year and Sales.

When you use the command `attach()` it is like saying in the office workspace (R), each one has a name assigned to him/her. But, it happens that both teams (data frames) have two members with the same name (Year).

```
Console Terminal Background Jobs
R 4.5.2 ~ /
> Marketing
# A tibble: 4 × 2
  Year FB_Followers
<dbl> <dbl>
1 2020 5000
2 2021 2000
3 2022 10000
4 2023 15000
> #The other is Sales and ut has two variables: Year and Sales in Dollars
> Sales
# A tibble: 4 × 2
  year Sales
<dbl> <dbl>
1 2010 5000
2 2011 2000
3 2012 3500
4 2013 5000
>
```

```
Console Terminal Background Jobs
R 4.5.2 ~ /
> #Now, let's use the command attach() for both data frames
> attach(Marketing)

The following objects are masked from Marketing (pos = 3):
  FB_Followers, Year

> attach(Sales)

The following object is masked _by_ .GlobalEnv:
  Sales

The following object is masked from Marketing (pos = 3):
  Year

The following object is masked from Marketing (pos = 4):
  Year
```

When you walk into the office (R) and call for team member (Year), you will get one of two responses:

1. one responds and the other does not (One masked the other) and you might discover you have speaking to the wrong person when it is too late in your conversation (modeling effort) This is called (**Masking Problem**)
2. none responds & people starts looking at you without saying anything (**an error message**)

```
> #Read the masking message !
> #Now, whenever you will call for (Year), it will only give you the entries
> #in the marketing ...Let's try
> Year
[1] 2010 2011 2012 2013
```

The right way of doing this is for you to walk into the office (R) and be specific about speaking with the team member (Year) who is part of the Marketing team (the Marketing data frame) We do this in R using the dollar sign (`Marketing$Year`).

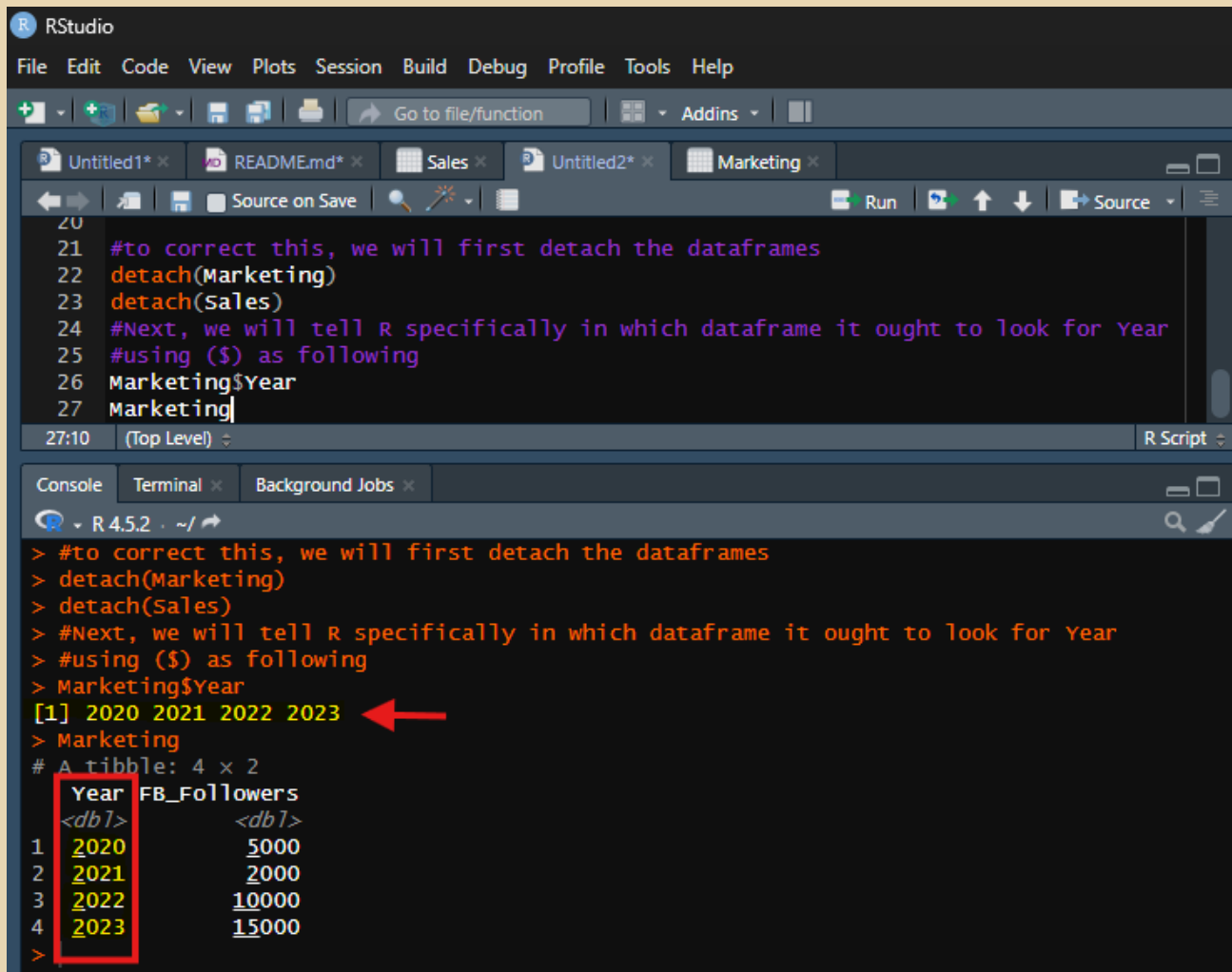
# The Data Dialogue.

*Simplified. Not Simplistic*

July. 29<sup>th</sup>. 2026 | Vol.1 | Issue 3 | Intro: Text Mining

In short, it is better not to attach your dataset from the very beginning.

But, if happens, you can correct this by using the command `detach()` and start calling the variables with their encompassing dataset as in the following screen.



The screenshot shows the RStudio interface. The script editor contains the following code:

```
20
21 #to correct this, we will first detach the dataframes
22 detach(Marketing)
23 detach(Sales)
24 #Next, we will tell R specifically in which dataframe it ought to look for Year
25 #using ($) as following
26 Marketing$Year
27 Marketing
```

The console shows the execution of these commands:

```
> #to correct this, we will first detach the dataframes
> detach(Marketing)
> detach(Sales)
> #Next, we will tell R specifically in which dataframe it ought to look for Year
> #using ($) as following
> Marketing$Year
[1] 2020 2021 2022 2023
> Marketing
# A tibble: 4 x 2
  Year FB_Followers
<dbl> <dbl>
1 2020         5000
2 2021         2000
3 2022        10000
4 2023        15000
```

A red arrow points to the output of `Marketing$Year`, and a red box highlights the `Marketing` tibble output.

# The Data Dialogue.

*Simplified. Not Simplistic*

July. 29<sup>th</sup>. 2026 | Vol.1 | Issue 3 | Intro: Text Mining

## *Pandora's Box Permitted or Plundered? The Ethics of Text Mining*

Text mining is part of data mining, which means extracting information from text available in different sources like archives, clinical and medical reports, amazon purchases, textbooks, photos, social media platforms' public posts and comments...etc. This provides massive information, mostly unstructured. With the help of programming codes, they bring such information into structure and extract patterns. An intense debate was sparked regarding the ethics of text mining.

***Arguments for text mining focus on efficiency and discovery.*** The technology saves thousands of hours of manual reading. It uncovers hidden patterns, trends, and correlations across millions of documents that humans could never synthesize manually.

***Conversely, arguments against text mining center on privacy, intellectual property, and data quality.*** The practice frequently scrapes data without explicit user consent, raising severe privacy concerns when personal information is involved. Copyright laws also present barriers, as mining proprietary databases or academic journals often leads to legal disputes over fair use. Algorithms struggle with sarcasm, cultural nuances, and context, often leading to inaccurate conclusions. Misinterpreting text can result in flawed business strategies or biased algorithmic decisions.

## *Before you Go*

Did it happen that you made a mistake while conducting data analysis? If so, send us a punchy 250 words describing this mistake and how you fixed it.

If it's humanized, we'll publish it next week with your name, title and link to your profile.

*Disclaimer: "We reserve the right to edit all submissions for conversational tone, length, and formatting to ensure it matches the Data Dialogue voice"*

You will find the submission button on my website.