

# Package: **disclosuR** (via r-universe)

September 16, 2024

**Type** Package

**Title** Text Conversion from Nexis Uni PDFs to R Data Frames

**Version** 0.5.2

**Date** 2023-06-05

**Description** Transform 'newswire' and earnings call transcripts as PDF obtained from 'Nexis Uni' to R data frames. Various 'newswires' and 'FairDisclosure' earnings call formats are supported. Further, users can apply several pre-defined dictionaries on the data based on Graffin et al. (2016)<[doi:10.5465/amj.2013.0288](https://doi.org/10.5465/amj.2013.0288)> and Gamache et al. (2015)<[doi:10.5465/amj.2013.0377](https://doi.org/10.5465/amj.2013.0377)>.

**License** GPL-3

**Imports** dplyr, lubridate, pdftools, qdap, SentimentAnalysis, stringi, stringr, syuzhet, zoo, SnowballC, tm, stats, rlang

**Encoding** UTF-8

**RoxygenNote** 7.2.3

**Repository** <https://jonasroettger.r-universe.dev>

**RemoteUrl** <https://github.com/jonasroettger/disclosur>

**RemoteRef** HEAD

**RemoteSha** 1ba128d76d92048ae75a41427668fa15ca7b4a77

## Contents

conference_call_segmenter . . . . .	2
conference_call_segmenter_folder . . . . .	3
impression_offsetting . . . . .	4
newswire_segmenter . . . . .	5
newswire_segmenter_folder . . . . .	6

<b>Index</b>	<b>8</b>
--------------	----------

---

conference\_call\_segmenter

*Earnings call segmenter*

---

## Description

Converts one earnings call transcript from 'FairDisclosure' obtained from 'NexisUni' to an R data frame.

## Usage

```
conference_call_segmenter(  
  file,  
  sentiment = FALSE,  
  emotion = FALSE,  
  regulatory_focus = FALSE,  
  laughter = FALSE,  
  narcissism = FALSE  
)
```

## Arguments

file	The name of the PDF file which the data are to be read from. If it does not contain an absolute path, the file name is relative to the current working directory, <code>getwd()</code> .
sentiment	Performs dictionary-based sentiment analysis based on the <a href="#">analyzeSentiment</a> function (default: FALSE)
emotion	Performs dictionary-based emotion analysis based on the <a href="#">get_nrc_sentiment</a> function (default: FALSE)
regulatory_focus	Calculates the number of words indicative for promotion and prevention focus based on the dictionary developed by <a href="#">Gamache et al., 2015</a> (default: FALSE)
laughter	Counts the number of times laughter was indicated in a quote. (default: FALSE)
narcissism	Counts the number of pronoun usage and calculates the ratio of first-person singular to first-person plural pronouns. This measure is derived from <a href="#">Zhu &amp; Chen, (2015)</a> (default: FALSE)

## Value

An R data frame with each row representing one quote. The columns indicate the quarter, year, section (presentation versus Q&A), the speaker's name, role, affiliation, and also three binary indicators on whether the speaker is the host company's (1) CEO, (2) CFO, and/or (3) Chairman.

## Examples

```
earnings_calls_df <- conference_call_segmenter(file = system.file("inst",
"examples",
"earnings_calls", "earnings_example_01.pdf",
package = "disclosuR"));
earnings_calls_df_sentiment <- conference_call_segmenter(file = system.file("inst",
"examples",
"newswire", "earnings_example_01.pdf",
package = "disclosuR"),
sentiment = TRUE);
```

---

conference\_call\_segmenter\_folder

*Earnings call segmenter (multiple files)*

---

## Description

Converts all 'FairDisclosure' earnings call transcripts obtained from 'NexisUni' in a folder to an R data frame.

## Usage

```
conference_call_segmenter_folder(
  folder_path,
  sentiment = FALSE,
  emotion = FALSE,
  regulatory_focus = FALSE,
  laughter = FALSE,
  narcissism = FALSE
)
```

## Arguments

folder_path	The name of the folder which the data are to be read from. If it does not contain an absolute path, the file name is relative to the current working directory.
sentiment	Performs dictionary-based sentiment analysis based on the <a href="#">analyzeSentiment</a> function (default: FALSE)
emotion	Performs dictionary-based emotion analysis based on the
regulatory_focus	Calculates the number of words indicative for promotion and prevention focus based on the dictionary developed by <a href="#">Gamache et al., 2015</a> (default: FALSE)
laughter	Counts the number of times laughter was indicated in a quote. (default: FALSE)
narcissism	Counts the number of pronoun usage and calculates the ratio of first-person singular to first-person plural pronouns. This measure is derived from <a href="#">Zhu &amp; Chen, (2015)</a> (default: FALSE)

## Value

An R data frame with each row representing one quote. The columns indicate the quarter, year, section (presentation versus Q&A), the speaker's name, role, affiliation, and also three binary indicators on whether the speaker is the host company's (1) CEO, (2) CFO, and/or (3) Chairman.

## Examples

```
earnings_calls_df <- conference_call_segementer_folder(  
  folder_path = system.file("inst",  
    "examples",  
    "earnings_calls",  
    package = "disclosuR");  
earnings_calls_df_sentiment <- conference_call_segementer_folder(  
  folder_path = system.file("inst",  
    "examples",  
    "newswire",  
    sentiment = TRUE,  
    package = "disclosuR");
```

---

impression\_offsetting *Impression offsetting*

---

## Description

Takes an event data set containing of dates and CUSIPs which have to correspond to a press data frame compiled by the function [newswire\\_segementer\\_folder](#).

## Usage

```
impression_offsetting(event_data, press_data_categorized)
```

## Arguments

event_data	An R data that contains three columns which have to be labeled "date_announced", "cusip", and "ID". The date_announced column contains the dates of the events for which impression offsetting is calculated. The cusip column contains the 8-digit cusip of the companies for which impression offsetting is calculated. The ID column should contain a unique ID that identifies the specific event.
press_data_categorized	An R data frame with each row representing one 'newswire' article. The columns indicate the title, text, 'newswire', date, and weekday. It should be the outcome of <a href="#">newswire_segementer</a> in which both the argument sentiment and text_clustering have been set to TRUE.

**Value**

An R data frame which contains the column of the event\_data plus three columns for the baseline announcements (positive, neutral, and negative) and three columns for the impression offsetting announcements (positive, neutral, and negative).

**Examples**

```
## Not run:
impression_offsetting(event_data, press_data)

## End(Not run)
```

---

newswire\_segmenter      *Newswire segmenter*

---

**Description**

Takes a PDF document containing a 'newswire' document obtained from 'NexisUni' and transforms it into an R data frame consisting of one row

**Usage**

```
newswire_segmenter(
  file,
  sentiment = FALSE,
  emotion = FALSE,
  regulatory_focus = FALSE,
  laughter = FALSE,
  narcissism = FALSE,
  text_clustering = FALSE
)
```

**Arguments**

file	The name of the PDF file which the data are to be read from. If it does not contain an absolute path, the file name is relative to the current working directory, <code>getwd()</code> .
sentiment	Performs dictionary-based sentiment analysis based on the <a href="#">analyzeSentiment</a> function (default: FALSE)
emotion	Performs dictionary-based emotion analysis based on the <a href="#">get_nrc_sentiment</a> function (default: FALSE)
regulatory_focus	Calculates the number of words indicative for promotion and prevention focus based on the dictionary developed by <a href="#">Gamache et al., 2015</a> (default: FALSE)
laughter	Counts the number of times laughter was indicated in a quote. (default: FALSE)

narcissism	Counts the number of pronoun usage and calculates the ratio of first-person singular to first-person plural pronouns. This measure is derived from <a href="#">Zhu &amp; Chen, (2015)</a> (default: FALSE)
text_clustering	Applies a document categorization using a dictionary developed based on the framework developed by <a href="#">Graffin et al., 2016</a> . (default: FALSE)

### Value

An R data frame with each row representing one 'newswire' article. The columns indicate the title, text, 'newswire', date, and weekday.

### Examples

```
newswire_df <- newswire_segementer(  
  file = system.file("inst",  
    "examples",  
    "newswire", "newswire_example_01.pdf",  
    package = "disclosuR");  
newswire_df_sentiment <- newswire_segementer(  
  file = system.file("inst",  
    "examples",  
    "newswire", "newswire_example_01.pdf",  
  sentiment = TRUE,  
  package = "disclosuR");
```

---

newswire\_segementer\_folder

*News wire segmenter (multiple files)*

---

### Description

Takes all PDF documents in a folder containing 'newswire' documents obtained from 'NexisUni' and transforms them into an R data frame consisting of one row per document.

### Usage

```
newswire_segementer_folder(  
  folder_path,  
  sentiment = FALSE,  
  emotion = FALSE,  
  regulatory_focus = FALSE,  
  laughter = FALSE,  
  narcissism = FALSE,  
  text_clustering = FALSE  
)
```

**Arguments**

folder_path	The path to the folder in which the 'newswire' PDFs reside. If it does not contain an absolute path, the folder name is relative to the current working directory, <code>getwd()</code> .
sentiment	Performs dictionary-based sentiment analysis based on the <a href="#">analyzeSentiment</a> function (default: FALSE)
emotion	Performs dictionary-based emotion analysis based on the <a href="#">get_nrc_sentiment</a> function (default: FALSE)
regulatory_focus	Calculates the number of words indicative for promotion and prevention focus based on the dictionary developed by <a href="#">Gamache et al., 2015</a> (default: FALSE)
laughter	Counts the number of times laughter was indicated in a quote. (default: FALSE)
narcissism	Counts the number of pronoun usage and calculates the ratio of first-person singular to first-person plural pronouns. This measure is derived from <a href="#">Zhu &amp; Chen, (2015)</a> (default: FALSE)
text_clustering	Applies a document categorization using a dictionary developed based on the framework developed by <a href="#">Graffin et al., 2016</a> . (default: FALSE)

**Value**

An R data frame with each row representing one 'newswire' article. The columns indicate the title, text, 'newswire', date, and weekday. (default: FALSE)

An R data frame with each row representing one 'newswire' article. The columns indicate the title, text, 'newswire', date, and weekday. Depending on the additional arguments, the output data can also contain sentiment, emotion, regulatory focus, laughter, narcissism and text cluster based on the Graffin et al. categories.

**Examples**

```
newswire_df <- newswire_segementer_folder(
  folder_path = system.file("inst",
    "examples",
    "newswire",
    package = "disclosuR"));
newswire_df_sentiment <- newswire_segementer_folder(
  folder_path = system.file("inst",
    "examples",
    "newswire",
    package = "disclosuR"), sentiment = TRUE);
```

# Index

`analyzeSentiment`, [2](#), [3](#), [5](#), [7](#)

`conference_call_segementer`, [2](#)

`conference_call_segementer_folder`, [3](#)

`get_nrc_sentiment`, [2](#), [5](#), [7](#)

`impression_offsetting`, [4](#)

`newswire_segementer`, [4](#), [5](#)

`newswire_segementer_folder`, [4](#), [6](#)