The R Package \texttt{sentometrics} to Compute, Aggregate and Predict with Textual Sentiment

David Ardia
Keven Bluteau
\textbf{Samuel Borms} (Université de Neuchâtel/Vrije Universiteit Brussel)
Kris Boudt
Text mining...

... is the process of distilling actionable insights from text.

Our focus is on textual sentiment analysis.
Time series econometrics...

... is the analysis of quantitative time series data typically in an economic context.

Our focus is on aggregation, econometric modelling and prediction.
The R package sentometrics

The package is a general framework that integrates (i) the qualification of sentiment from texts, (ii) the aggregation into different sentiment measures, and (iii) the optimized prediction based on these measures.

Typical workflow:

**STEP 1**
Build a corpus of texts with quantifiable metadata (“features”)

**STEP 2**
Pick lexicons and compute textual sentiment

**STEP 3**
Aggregate document-level sentiment scores into time series

**STEP 4**
Estimate a sentiment-based prediction model

**STEP 5**
Evaluate model performance and sentiment attribution
Let’s go for a run with sentometrics

```r
library("sentometrics")

We have a built-in dataset of news articles between 1995 and 2014, from The Wall Street Journal and The Washington Post.

data("usnews", package = "sentometrics")
```

<table>
<thead>
<tr>
<th>ID</th>
<th>DATE</th>
<th>TEXT</th>
<th>WSJ</th>
<th>WAPO</th>
<th>ECONOMY</th>
<th>NONECONOMY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1995-01-02</td>
<td>Full text 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1995-01-05</td>
<td>Full text 2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Features:** relevance/importance indicators & selectors.
**Massage the corpus**

Checking the requirements of the corpus.

```r
corpusAll <- sento_corpus(usnews)
```

Subsetting the corpus, using the `quanteda` package.

```r
corpus <- quanteda::corpus_subset(corpusAll, date < "2014-10-01")
```

Adding features (for example: entities, topics, events).

```r
regex <- c("\bRepublic[s]?\b|\bDemocrat[s]?\b|\belection\b|\b[US|U.S.][pP]resident\b|\bwar\b")
corpus <- add_features(corpus,
    keywords = list(uncert = "uncertainty",
                    uselect = regex,
                    finance = c("\bstock market")),
    do.binary = TRUE,
    do.regex = c(FALSE, TRUE, TRUE))
```
Pick the word lists for lexicon-based sentiment analysis

We have English, Dutch and French built-in word lists.

```r
data("list_lexicons", package = "sentometrics")
data("list_valence_shifters", package = "sentometrics")
```

Prepare and check the lexicons.

```r
lex <- setup_lexicons(lexiconsIn = list_lexicons[c("LM_en", "HENRY_en")],
valenceIn = list_valence_shifters[["en"]])
```
From sentiment to time series: aggregation specs

Aggregation of the many sentiment scores...
  ... within documents = document-level sentiment
  ... across documents = time series
  ... across time = smoothed time series

... across lexicons, features and time aggregation schemes

One control function to define all of this.

```
ctrAgg <- ctr_agg(howWithin = "tf-idf",
                 howDocs = "proportional",
                 howTime = c("equal_weight", "linear", "almon"),
                 do.ignoreZeros = TRUE,
                 by = "month",
                 fill = "zero",
                 lag = 12,
                 ordersAlm = 1:3,
                 do.inverseAlm = TRUE)
```

STEPS 2–3
Create many sentiment time series at once

This one simple function call gives you a wide number of different sentiment time series, or “measures”.

```
sentMeas <- sento_measures(corpus, lexicons = lex, ctr = ctrAgg)
```

The sentiment measures are represented as “lexicon—feature—smoothing”.

```
head(sentMeas[["measures"]][, 1:5])
```

- **Lexicon**: The specific lexicon used for sentiment analysis.
- **Feature**: The feature or category of sentiment analysis.
- **Time Aggregation Scheme**: The method used to aggregate sentiment over time.

**Steps 2-3**
Plot across a given time series dimension

```r
plot(measures_select(sentMeas,
    c("economy", "finance", "uncert"),
    do.combine = FALSE),
group = "features")
```
We try to predict the monthly U.S. EPU index...

The Economic Policy Uncertainty (EPU) index is a partly news-based measure of policy-related economic uncertainty. It is served with the package as a dataset.

http://www.policyuncertainty.com
We propose to use the elastic net regression (relying on glmnet), which balances between the LASSO and Ridge regressions through an $\alpha$ parameter. The large number and collinearity of the sentiment measures motivate this choice. A straightforward control function defines the model setup.

$$y_{u+h} = \delta + \gamma^T x_u + \beta_1 s_{u1}^1 + \ldots + \beta_p s_{up}^p + \ldots + \beta_p s_{up}^p + \epsilon_{u+h}$$

A straightforward control function defines the model setup.

```r
ctrIter <- ctr_model(model = "gaussian",
                     type = "BIC",
                     h = 1,
                     alphas = c(0.3, 0.5, 0.7),
                     do.iter = TRUE,
                     nSample = 36)
```

STEPS 4–5
Run the prediction model iteratively

Load the data.
```r
data("epu", package = "sentometrics")
y <- epu[epu["date"] >= sentMeas[["measures"]][["date"]][1], "index"]
```

Running the out-of-sample prediction analysis is easy.
```r
out <- sento_model(sentMeas, y, ctr = ctrIter)
```

We call “attribution” the decomposition of the prediction into one of the underlying sentiment time series dimensions.
```r
attr <- retrieve_attributions(out, sentMeas, do.normalize = FALSE)
```
Visualise the out-of-sample prediction and attribution

plot(out)

plot_attrictions(attr, group = "features")
The package offers considerable flexibility to develop textual sentiment time series.

Many potential uses of the framework in financial and investment analysis:

- Following up of macro-trends
- Stock screening across a set of pre-defined features (e.g. sustainability)
- Abnormal sentiment detection
- Sentiment-based trading strategy
- ...

Go out there and test the package!