Textifier

Summary

- Textifier, a text classification software.
- Combines statistical classifiers (based on open source WEKA toolkit) with classification based on keyword counting.
- Keyword counting helps for stability and with sparse data situations.

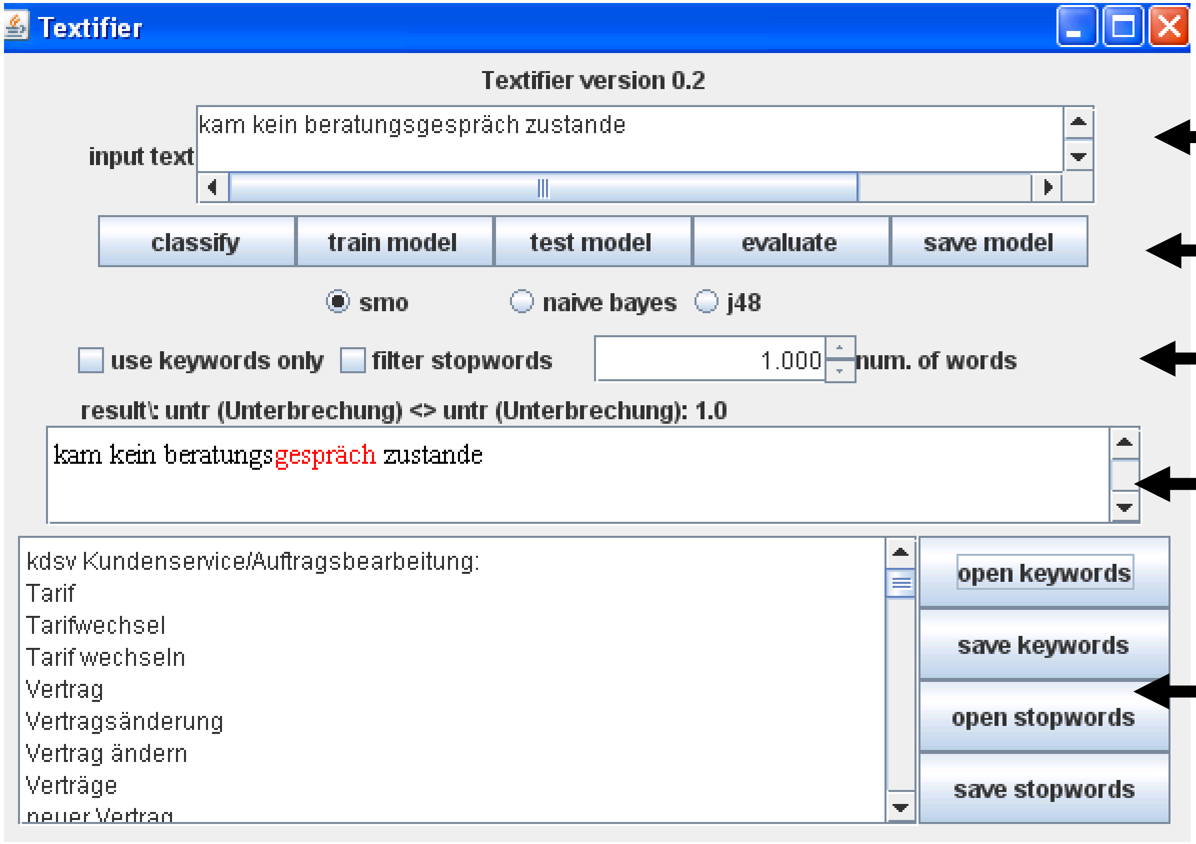
Textifier

Overview

- Assigns text to categories
- Written in Java
- Uses two interfaces:
  - Webservice / XML format
  - GUI Application as administration interface
- Categorizes
  - by WEKA classifiers
  - by keyword counting

Textifier

Administration interface



Textifier version 0.2

input text: kam kein beratungsgespräch zustande

● smo ○ naive bayes ○ j48

□ use keywords only □ filter stopwords 1.000 num. of words

result\: untr (Unterbrechung) <> untr (Unterbrechung): 1.0

kam kein beratungsgespräch zustande

kdsv Kundenservice/Auftragsbearbeitung:
Tarif
Tarifwechsel
Tarif wechseln
Vertrag
Vertragsänderung
Vertrag ändern
Verträge
neuer Vertrag

Input

Manage classifier models

Filter training material

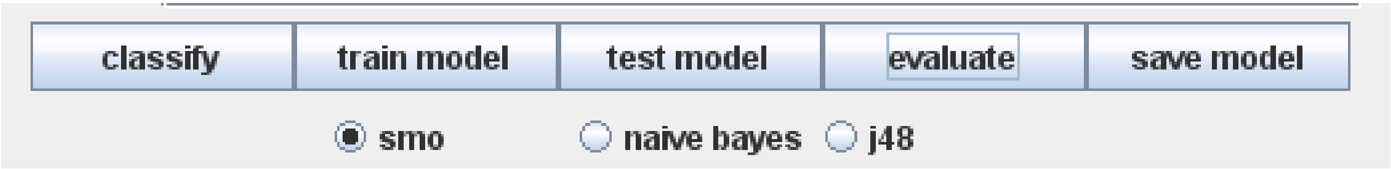
Output

Edit configuration

Textifier

Manage classifier models

- Generate, test, save WEKA model to file
- Classifier types: SVM, Bayes, Tree

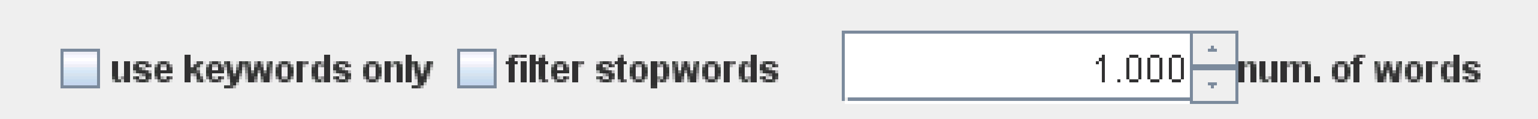


● smo ○ naive bayes ○ j48

Textifier

Filter training text

- For construction of word vectors from text:
- Use only keywords
- Filter stop words
- Number of words to use

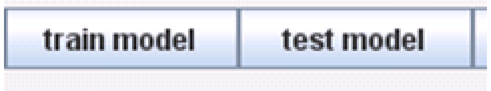


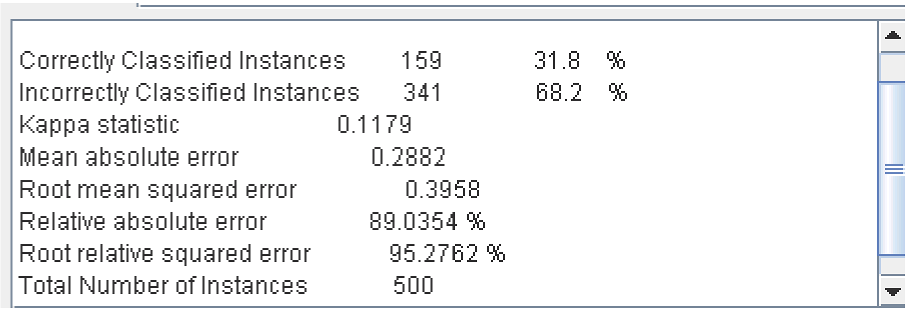
□ use keywords only □ filter stopwords 1.000 num. of words

Textifier

Train/test models

- Generate WEKA models with different configurations and evaluate them.



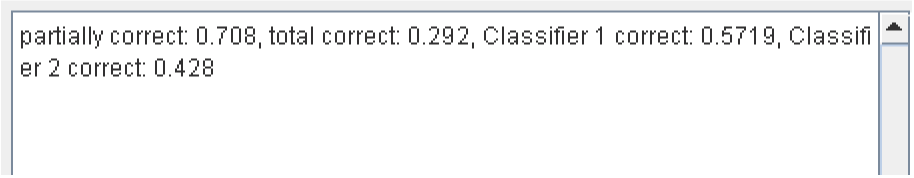


```
Correctly Classified Instances       159           31.8   %
Incorrectly Classified Instances     341           68.2   %
Kappa statistic                   0.1179
Mean absolute error                  0.2882
Root mean squared error               0.3958
Relative absolute error              89.0354 %
Root relative squared error          95.2762 %
Total Number of Instances            500
```

Textifier

Evaluate classification.

- Classify test set with statistical AND keyword-based classifier.

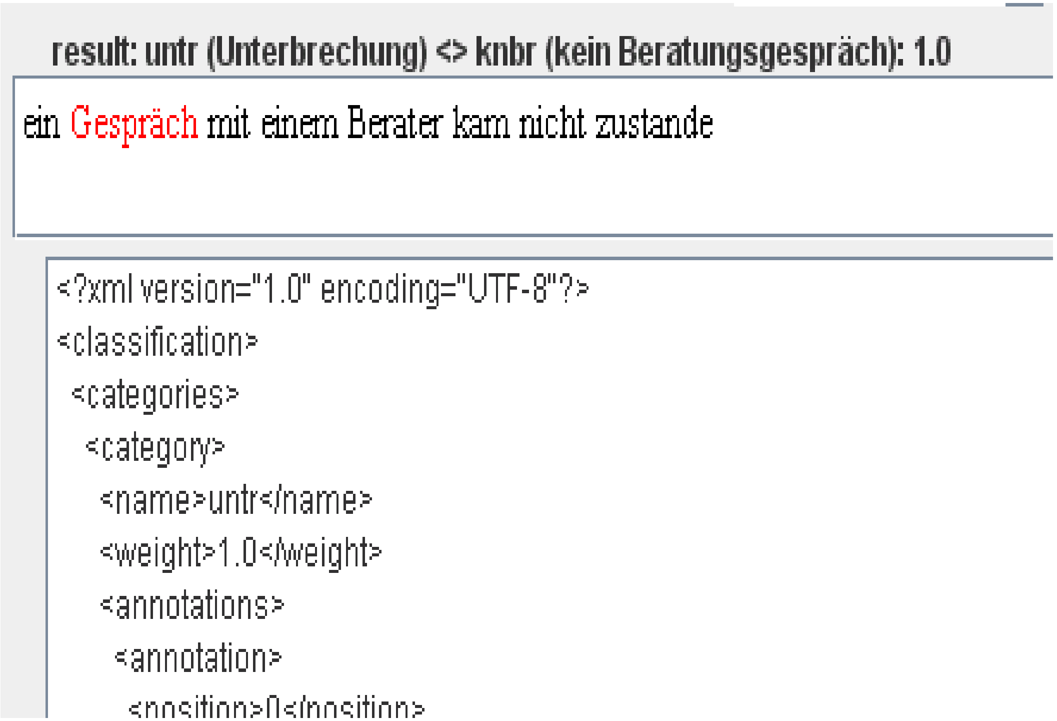


```
partially correct: 0.708, total correct: 0.292, Classifier 1 correct: 0.5719, Classifi
er 2 correct: 0.428
```

Textifier

Ouput

- Winner categories from two classifiers
- Markup of content words
- XML formatted output



result: untr (Unterbrechung) <> knbr (kein Beratungsgespräch): 1.0

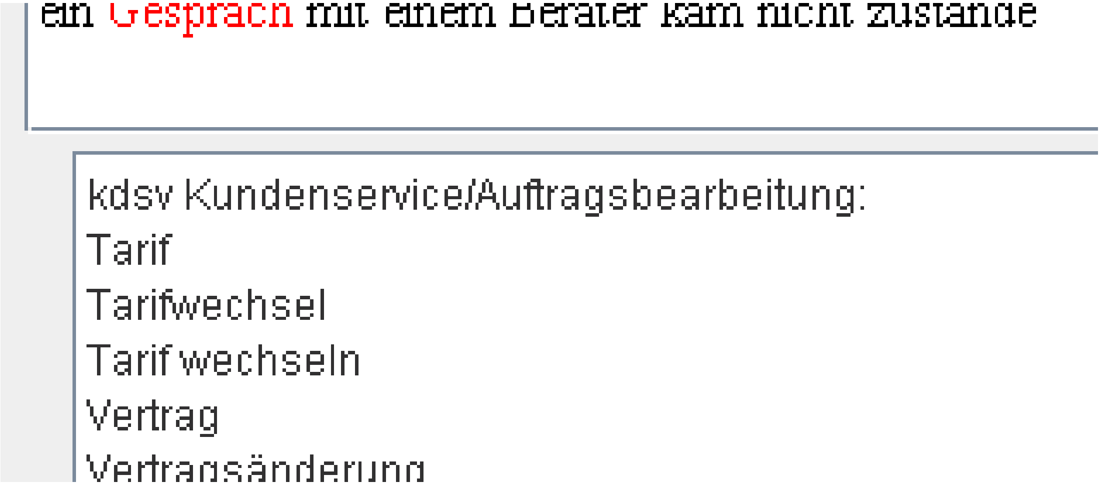
ein Gespräch mit einem Berater kam nicht zustande

```
<?xml version="1.0" encoding="UTF-8"?>
<classification>
 <categories>
  <category>
   <name>untr</name>
   <weight>1.0</weight>
   <annotations>
    <annotation>
     <position>0</position>
```

Textifier

Category definition

- Entry in keyword configuration
- Format <ID: single token> <name: 1-n token> :



ein Gespräch mit einem Berater kam nicht zustande

kdsv Kundenservice/Auftragsbearbeitung:
Tarif
Tarifwechsel
Tarif wechseln
Vertrag
Vertragsänderung

Textifier

Webservice interface

- XML format

```
<?xml version="1.0" encoding="UTF-8"?>
<classification>
  <categories>
    <category>
      <name>untr</name>
      <weight>1.0</weight>
      <annotations>
        <annotation>
          <position>0</position>
          <token>
            <name>gespräch</name>
            <gramsize>1</gramsize>
            <weight>1.0</weight>
          </token>
          <string>gespräch</string>
        </annotation>
      </annotations>
    </category>
    <category>
      <name>knbr</name>
      <weight>1.0</weight>
      <annotations>
        <annotation>
          <position>0</position>
          <token>
            <name>gespräch</name>
            <gramsize>1</gramsize>
            <weight>1.0</weight>
          </token>
          <string>gespräch</string>
        </annotation>
      </annotations>
    </category>
  </categories>
  <text>ein Gespräch mit einem Berater kam nicht zustande</text>
</classification>
```