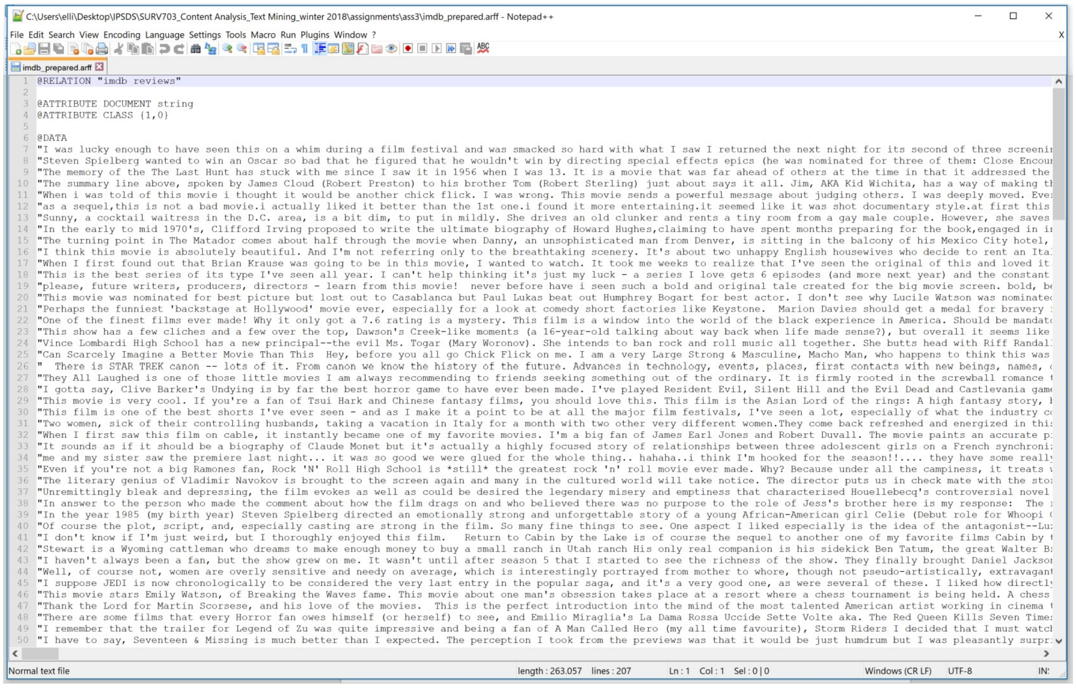
SURV703 – Content Analysis
“Testing” WEKA, explaining the results
Elisabeth Linek

Task: Install the appropriate version of WEKA (Links zu einer externen Webseite.) for your system. Convert the dataset into the ARFF format required by WEKA. Preprocess it with settings of your choice in the WEKA Explorer. Finally, perform a binary sentence classification (as shown in this week's lesson) with algorithms of your choice in the WEKA Experimenter. Hand in a table with the results and a brief interpretation in textual form.

First of all I followed the examples from the video lecture in order to have a “guided walk” through the WEKA GUI.

Starting with the preparation of the csv-file, in order get it into a WEKA-acceptable format, I set up the following:


C:\Users\elli\Desktop\IPSDS\SURV703_Content Analysis_Text Mining_winter 2018\assignments\ass3\imdb_prepared.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
imdb_prepared.arff

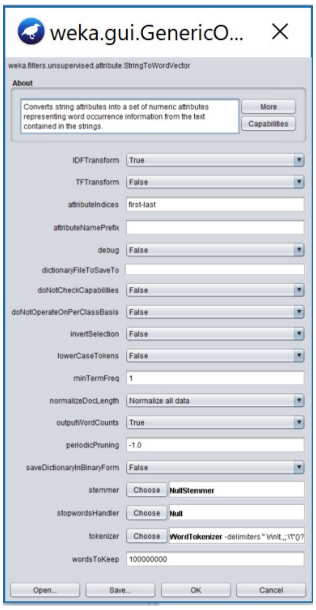
```
@RELATION "imdb reviews"

@ATTRIBUTE DOCUMENT string
@ATTRIBUTE CLASS {1,0}

@DATA
"I was lucky enough to have seen this on a whim during a film festival and was smacked so hard with what I saw I returned the next night for its second of three screeni
"Steven Spielberg wanted to win an Oscar so bad that he figured that he wouldn't win by directing special effects epics (he was nominated for three of them: Close Encou
"The memory of the The Last Hunt has stuck with me since I saw it in 1956 when I was 13. It is a movie that was far ahead of others at the time in that it addressed the
"The summary line above, spoken by James Cloud (Robert Preston) to his brother Tom (Robert Sterling) just about says it all. Jim, AKA Kid Wichita, has a way of making t
"When i was told of this movie i thought it would be another chick flick. I was wrong. This movie sends a powerful message about judging others. I was deeply moved. Eve
"as a sequel,this is not a bad movie.i actually liked it better than the 1st one.i found it more entertaining.it seemed like it was shot documentary style.at first this
"Sunny, a cocktail waitress in the D.C. area, is a bit dim, to put in mildly. She drives an old clunker and rents a tiny room from a gay male couple. However, she saves
"In the early to mid 1970's, Clifford Irving proposed to write the ultimate biography of Howard Hughes,claiming to have spent months preparing for the book,engaged in i
"The turning point in The Matador comes about half through the movie when Danny, an unsophisticated man from Denver, is sitting in the balcony of his Mexico City hotel,
"I think this movie is absolutely beautiful. And I'm not referring only to the breathtaking scenery. It's about two unhappy English housewives who decide to rent an Ita
"When I first found out that Brian Krause was going to be in this movie, I wanted to watch. It took me weeks to realize that I've seen the original of this and loved it
"This is the best series of its type I've seen all year. I can't help thinking it's just my luck - a series I love gets 6 episodes (and more next year) and the constant
"please, future writers, producers, directors - learn from this movie!  never before have i seen such a bold and original tale created for the big movie screen. bold, b
"This movie was nominated for best picture but lost out to Casablanca but Paul Lukas beat out Humphrey Bogart for best actor. I don't see why Lucile Watson was nominate
"Perhaps the funniest 'backstage at Hollywood' movie ever, especially for a look at comedy short factories like Keystone.  Marion Davies should get a medal for bravery
"One of the finest films ever made! Why it only got a 7.6 rating is a mystery. This film is a window into the world of the black experience in America. Should be mandat
"This show has a few cliches and a few over the top, Dawson's Creek-like moments (a 16-year-old talking about way back when life made sense?), but overall it seems like
"Vince Lombardi High School has a new principal--the evil Ms. Togar (Mary Woronov). She intends to ban rock and roll music all together. She butts head with Riff Randal
"Can Scarcely Imagine a Better Movie Than This  Hey, before you all go Chick Flick on me. I am a very Large Strong & Masculine, Macho Man, who happens to think this was
"  There is STAR TREK canon -- lots of it. From canon we know the history of the future. Advances in technology, events, places, first contacts with new beings, names,
"They All Laughed is one of those little movies I am always recommending to friends seeking something out of the ordinary. It is firmly rooted in the screwball romance
"I gotta say, Clive Barker's Undying is by far the best horror game to have ever been made. I've played Resident Evil, Silent Hill and the Evil Dead and Castlevania gam
"This movie is very cool. If you're a fan of Tsui Hark and Chinese fantasy films, you should love this. This film is the Asian Lord of the rings: A high fantasy story,
"This film is one of the best shorts I've ever seen - and as I make it a point to be at all the major film festivals, I've seen a lot, especially of what the industry c
"Two women, sick of their controlling husbands, taking a vacation in Italy for a month with two other very different women.They come back refreshed and energized in thi
"When I first saw this film on cable, it instantly became one of my favorite movies. I'm a big fan of James Earl Jones and Robert Duvall. The movie paints an accurate p
"It sounds as if it should be a biography of Claude Monet but it's actually a highly focused story of relationships between three adolescent girls on a French synchroni
"me and my sister saw the premiere last night... it was so good we were glued for the whole thing.. hahaha..i think I'm hooked for the season!!.... they have some reall
"Even if you're not a big Ramones fan, Rock 'N' Roll High School is *still* the greatest rock 'n' roll movie ever made. Why? Because under all the campiness, it treats
"The literary genius of Vladimir Navokov is brought to the screen again and many in the cultured world will take notice. The director puts us in check mate with the sto
"Unremittingly bleak and depressing, the film evokes as well as could be desired the legendary misery and emptiness that characterised Houellebecq's controversial novel
"In answer to the person who made the comment about how the film drags on and who believed there was no purpose to the role of Jess's brother here is my response:  The
"In the year 1985 (my birth year) Steven Spielberg directed an emotionally strong and unforgettable story of a young African-American girl Celie (Debut role for Whoopi
"Of course the plot, script, and, especially casting are strong in the film. So many fine things to see. One aspect I liked especially is the idea of the antagonist--Lu
"I don't know if I'm just weird, but I thoroughly enjoyed this film.   Return to Cabin by the Lake is of course the sequel to another one of my favorite films Cabin by
"Stewart is a Wyoming cattleman who dreams to make enough money to buy a small ranch in Utah ranch His only real companion is his sidekick Ben Tatum, the great Walter B
"I haven't always been a fan, but the show grew on me. It wasn't until after season 5 that I started to see the richness of the show. They finally brought Daniel Jackso
"Well, of course not, women are overly sensitive and needy on average, which is interestingly portrayed from mother to whore, though not pseudo-artistically, extravagan
"I suppose JEDI is now chronologically to be considered the very last entry in the popular saga, and it's a very good one, as were several of these. I liked how directl
"This movie stars Emily Watson, of Breaking the Waves fame. This movie about one man's obsession takes place at a resort where a chess tournament is being held. A chess
"Thank the Lord for Martin Scorsese, and his love of the movies.  This is the perfect introduction into the mind of the most talented American artist working in cinema
"There are some films that every Horror fan owes himself (or herself) to see, and Emilio Miraglia's La Dama Rossa Uccide Sette Volte aka. The Red Queen Kills Seven Time
"I remember that the trailer for Legend of Zu was quite impressive and being a fan of A Man Called Hero (my all time favourite), Storm Riders I decided that I must watc
"I have to say, Seventeen & Missing is much better than I expected. The perception I took from the previews was that it would be just humdrum but I was pleasantly surpr
```

Normal text file length : 263.057 lines : 207 Ln : 1 Col : 1 Sel : 0 | 0 Windows (CR LF) UTF-8 IN!

I saved the text-file with the extension *.arff (imdb_bow.arff)*, in order to import it as data file into the WEKA GUI. After loading the data file into WEKA, I followed the explained filter adoptions, changing the preselection into the following:



weka.filters.unsupervised.attribute.StringToWordVector

Converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings.

IDFTransform	True
TFTransform	False
attributeIndices	first-last
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	False
minTermFreq	1
normalizeDocLength	Normalize all data
outputWordCounts	True
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	Choose **NullStemmer**
stopwordsHandler	Choose **Null**
tokenizer	Choose **WordTokenizer** -delimiters " \r\n\t.,;:\'\"()?
wordsToKeep	100000000

After the pre-selection was changed, the filter was applied to the DOCUMENT, and I went on to define or set the CLASS argument as attribute for the planned "implementation" of algorithms.
As final step of the preparation I saved the vectorized and normalized data set, giving it the name.
imdb_prepared_bow.arff

_____

Now starting with the application of the algorithms, I will jump to discuss briefly the results.

As shown within the video lectures I went on with a J48-tree algorithm. I changed the percentage of trainign data first to 75% and in a second step to 80%, just to find out what impact such a change might have on the results:

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
[...]
Instances:    200
Attributes:   8153
Test mode:    split 75.0% train, remainder test

J48 pruned tree
------------------
awful <= 0
|   worst <= 1.967162
|   |   entire <= 1.022614
|   |   |   nothing <= 1.83166
|   |   |   |   20 <= 1.056217
|   |   |   |   |   John <= 0
|   |   |   |   |   |   Some <= 0
|   |   |   |   |   |   |   family <= 1.554192
|   |   |   |   |   |   |   |   & <= 0
|   |   |   |   |   |   |   |   |   2 <= 0
|   |   |   |   |   |   |   |   |   |   away <= 2.129845
|   |   |   |   |   |   |   |   |   |   |   used <= 0
|   |   |   |   |   |   |   |   |   |   |   |   before <= 2.105742
|   |   |   |   |   |   |   |   |   |   |   |   |   minutes <= 1.992505
|   |   |   |   |   |   |   |   |   |   |   |   |   |   however <= 1.268507
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   and <= 0.060602
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   a <= 0.473168: 0 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   a > 0.473168: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   and > 0.060602
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   were <= 2.622447
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   comedy <= 2.572386: 1 (68.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   comedy > 2.572386
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   It <= 1.069835: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   It > 1.069835: 0 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   were > 2.622447
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   This <= 0.718081: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   This > 0.718081: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   however > 1.268507
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   But <= 0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   But > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   minutes > 1.992505: 0 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   before > 2.105742: 0 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   used > 0
|   |   |   |   |   |   |   |   |   |   |   |   However <= 0: 0 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   However > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   away > 2.129845: 0 (6.0/1.0)
|   |   |   |   |   |   |   |   |   2 > 0: 0 (6.0/1.0)
|   |   |   |   |   |   |   |   & > 0: 1 (7.0)
|   |   |   |   |   |   |   family > 1.554192: 0 (8.0/1.0)
|   |   |   |   |   |   Some > 0: 0 (7.0/1.0)
|   |   |   |   |   John > 0: 1 (8.0)
|   |   |   |   20 > 1.056217: 0 (8.0/1.0)
|   |   |   nothing > 1.83166: 0 (10.0)
|   |   entire > 1.022614: 0 (9.0)
|   worst > 1.967162: 0 (11.0)
awful > 0: 0 (14.0)

Number of Leaves: 24
Size of the tree:      47

=== Summary ===
Correctly Classified Instances        27               54      %
Incorrectly Classified Instances      23               46      %
Kappa statistic                        0.077
Mean absolute error                    0.4573
Root mean squared error                0.6606
Relative absolute error               91.2745 %
Root relative squared error          131.7917 %
Total Number of Instances             50

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
              0,522    0,444    0,500      0,522   0,511      0,077   0,553     0,491     1
              0,556    0,478    0,577      0,556   0,566      0,077   0,553     0,575     0
Weighted
Avg.          0,540    0,463    0,542      0,540   0,541      0,077   0,553     0,536
```

```
=== Confusion Matrix ===
 a  b   <-- classified as
 12 11 |  a = 1
 12 15 |  b = 0
```

Followed by a J48 tree algorithm based on 80% of trainign data from the data set:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
[...]
Instances:    200
Attributes:   8153
Test mode:    split 80.0% train, remainder test

J48 pruned tree
------------------
awful <= 0
|   worst <= 1.967162
|   |   entire <= 1.022614
|   |   |   nothing <= 1.83166
|   |   |   |   20 <= 1.056217
|   |   |   |   |   John <= 0
|   |   |   |   |   |   Some <= 0
|   |   |   |   |   |   |   family <= 1.554192
|   |   |   |   |   |   |   |   & <= 0
|   |   |   |   |   |   |   |   |   2 <= 0
|   |   |   |   |   |   |   |   |   |   away <= 2.129845
|   |   |   |   |   |   |   |   |   |   |   used <= 0
|   |   |   |   |   |   |   |   |   |   |   |   before <= 2.105742
|   |   |   |   |   |   |   |   |   |   |   |   |   minutes <= 1.992505
|   |   |   |   |   |   |   |   |   |   |   |   |   |   however <= 1.268507
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   and <= 0.060602
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   a <= 0.473168: 0 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   a > 0.473168: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   and > 0.060602
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   were <= 2.622447
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   comedy <= 2.572386: 1 (68.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   comedy > 2.572386
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   It <= 1.069835: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   It > 1.069835: 0 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   were > 2.622447
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   This <= 0.718081: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   This > 0.718081: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   however > 1.268507
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   But <= 0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   But > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   minutes > 1.992505: 0 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   before > 2.105742: 0 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   used > 0
|   |   |   |   |   |   |   |   |   |   |   |   However <= 0: 0 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   However > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   away > 2.129845: 0 (6.0/1.0)
|   |   |   |   |   |   |   |   |   2 > 0: 0 (6.0/1.0)
|   |   |   |   |   |   |   |   & > 0: 1 (7.0)
|   |   |   |   |   |   |   family > 1.554192: 0 (8.0/1.0)
|   |   |   |   |   |   Some > 0: 0 (7.0/1.0)
|   |   |   |   |   John > 0: 1 (8.0)
|   |   |   |   20 > 1.056217: 0 (8.0/1.0)
|   |   |   nothing > 1.83166: 0 (10.0)
|   |   entire > 1.022614: 0 (9.0)
|   worst > 1.967162: 0 (11.0)
awful > 0: 0 (14.0)

Number of Leaves: 24
Size of the tree:       47

=== Summary ===
Correctly Classified Instances          20               50      %
Incorrectly Classified Instances        20               50      %
Kappa statistic                          0.0123
Mean absolute error                      0.5074
Root mean squared error                  0.6868
Relative absolute error                101.4189 %
Root relative squared error            137.2681 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,632    0,619    0,480      0,632    0,545      0,013    0,514     0,499     1
                 0,381    0,368    0,533      0,381    0,444      0,013    0,514     0,521     0
Weighted
Avg.             0,500    0,487    0,508      0,500    0,492      0,013    0,514     0,511

=== Confusion Matrix ===

 a  b   <-- classified as
 12  7 |  a = 1
 13  8 |  b = 0
```

→ I have to admit, that I was surprised by the difference, just based on the change of the amount of training data compared to the remaining test data. The correctly classified instances differ from 54% to 50%. The confusion matrix differs as well.

What surprised me most is the fact, that the level of precision differs much from the example we saw on the video lectures, where we reached levels around 90%. That will depend on the provided data set, but I would like to know which other filters I could have set to reach better results.

As another test I went on with a Naive Bayes algorithm:

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
[...]
Instances:    200
Attributes:   8153
Test mode:    split 80.0% train, remainder test
```

Naive Bayes Classifier (just some examples, the whole result overview would have needed much more space)

```
Attribute                  Class
                          1      0
                        (0.5)  (0.5)
==============================================================
$1
  mean                   0.0187      0
  std. dev.              0.3113 0.3113
  weight sum                100    100
  precision               1.868  1.868

&
  mean                   0.6253 0.0853
  std. dev.              2.3977 0.6298
  weight sum                100    100
  precision              1.4212 1.4212

*
  mean                   0.0374      0
  std. dev.              0.6227 0.6227
  weight sum                100    100
  precision              3.7361 3.7361

Jack
  mean                   0.0423 0.1479
  std. dev.              0.3521 1.0877
  weight sum                100    100
  precision              2.1125 2.1125

friend
  mean                   0.1225 0.2205
  std. dev.              0.5084 1.0948
  weight sum                100    100
  precision              0.8167 0.8167

friends
  mean                    0.217 0.1346
  std. dev.              0.8292 0.6458
  weight sum                100    100
  precision               0.434  0.434

frightening
  mean                   0.0803 0.0357
  std. dev.              0.4567 0.3552
  weight sum                100    100
  precision              0.8924 0.8924

yes
  mean                        0 0.0887
  std. dev.              0.2956 0.5814
  weight sum                100    100
  precision              1.7734 1.7734

zombie/cannibal
  mean                        0 0.0764
  std. dev.              1.2741 1.2741
  weight sum                100    100
  precision              7.6448 7.6448
```

```
=== Summary ===

Correctly Classified Instances          22               55      %
Incorrectly Classified Instances        18               45      %
Kappa statistic                          0.0932
Mean absolute error                      0.4537
Root mean squared error                  0.6712
Relative absolute error                 90.6865 %
Root relative squared error            134.1532 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,474    0,381    0,529      0,474    0,500      0,094    0,590     0,539     1
               0,619    0,526    0,565      0,619    0,591      0,094    0,564     0,552     0
```

```
Weighted
Avg.          0,550     0,457     0,548     0,550    0,548     0,094     0,576     0,546

=== Confusion Matrix ===

  a  b   <-- classified as
  9 10 |  a = 1
  8 13 |  b = 0
```

The results highlighted here are close to the results I reached with the J48 algorithm, based on 75% training data. The confusion matrix showing, just as the correction level already did, that the algorithm are not leading to a very successful automatic classification. I am still wondering what I depends on. Maybe the normalisation of the original data was not fitting well?

Leaving the classification based WEKA GUI behind, turning to the experimental tool, in order to compare the algorithms, I got the following results.
In a first step I selected the following algorithms which lead to an overload of m system, and I interrupted the test.

- ZeroR: No model built yet.
- Naive Bayes Classifier: No model built yet.
- SMO: No model built yet.
- IBk: No model built yet.
- J48
- Decision Stump: No model built yet. (left aside in the second trial)

I decided to work according to the example from lecture: 10 repetitions each based on 10% of the data.

The results:

```
18:03:21: Started
20:14:03: User aborting experiment.
20:14:51: Interrupted
20:14:51: There were 0 errors
```

I went on with a second try, downgrading the number of folds to 5 (even though it is not the prefered method), the second trial was successful.

1) Percent of correctness selected: (according to video lectures)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix [...]
Analysing:  Percent_correct
Datasets:    1
Resultsets: 5
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       13.12.18 20:32

Dataset                          (1) rules.Ze | (2) bayes (3) funct (4) lazy.   (5) trees
-------------------------------------------------------------------------------------------
'imdb reviews-weka.filter (50)   50.00 |   65.25 v       74.55 v   59.35 v   62.05 v
-------------------------------------------------------------------------------------------
                                    (v/ /*) |   (1/0/0)   (1/0/0)   (1/0/0)   (1/0/0)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) bayes.NaiveBayes '' 5995231201785697655
(3) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -
calibrator \"functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"' -6585883636378691736
(4) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -
3080186098777067172
(5) trees.J48 '-C 0.25 -M 2' -217733168393644444
```

→ The comparison of the given levels of correct classifications show, that the “function SMO” algorithm reached the highest level.

2) Precision selected: (as shown in video)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
"weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 2 -
stddev-width 2 -sig-width 1 -count-width 5 -print-col-names -print-row-names -enum-col-names"
Analysing:  IR_precision
Datasets:   1
Resultsets: 5
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       13.12.18 20:35


Dataset                   (1) rules.Z | (2) baye (3) func (4) lazy (5) tree
----------------------------------------------------------------------------
'imdb reviews-weka.filter (50)   0.50 |   0.68 v    0.73 v    0.58 v   0.63 v
----------------------------------------------------------------------------
                              (v/ /*) |  (1/0/0)   (1/0/0)   (1/0/0)  (1/0/0)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) bayes.NaiveBayes '' 5995231201785697655
(3) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -
calibrator \"functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"' -6585883636378691736
(4) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -
3080186098777067172
(5) trees.J48 '-C 0.25 -M 2' -217733168393644444
```

→ The comparison of precision levels shows as well, that the “function SMO” algorithm reaches the highest level of correct classifications. But still, we do not reach a level of 80% or higher, as we saw in the video lectures.

3) F measure selected: (as shown in video lectures)

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
"weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 2 -
stddev-width 2 -sig-width 1 -count-width 5 -print-col-names -print-row-names -enum-col-names"
Analysing:  F_measure
Datasets:   1
Resultsets: 5
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       13.12.18 20:36


Dataset                   (1) rules.Z | (2) baye (3) func (4) lazy (5) tree
----------------------------------------------------------------------------
'imdb reviews-weka.filter (50)   0.67 |   0.62      0.75 v    0.65     0.61
----------------------------------------------------------------------------
                              (v/ /*) |  (0/1/0)  (1/0/0)   (0/1/0)  (0/1/0)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) bayes.NaiveBayes '' 5995231201785697655
(3) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -
calibrator \"functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"' -6585883636378691736
(4) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -
3080186098777067172
(5) trees.J48 '-C 0.25 -M 2' -217733168393644444
```

Overview of analysis:

- 20:32:13 - Available resultsets
- 20:32:48 - Percent_correct - rules.ZeroR '' 48055541465867954
- 20:35:21 - IR_precision - rules.ZeroR '' 48055541465867954
- 20:36:44 - F_measure - rules.ZeroR '' 48055541465867954

⟶ The comparison of the f test levels shows what we could see before, the“function SMO” algorithm reaches the highest level regarding the classifications.
But still, we do not reach the level we saw in the video lectures.

I would assume, that further data preparation at the beginning of the planned classification could lead to higher or better results, with less confusion within the classifications.