langnet

Toward a Complex Networks Approach on Text Type Classification

D. Margan, A. Meštrović, M. Ivašić-Kos, S. Martinčić-Ipšić

Department of Informatics, University of Rijeka

{dmargan,amestrovic,marinai,smarti}@uniri.hr

Introduction and motivation

- The growing amount of text electronically available has placed text type classification among essential issues in the field of text mining

- Text type classification by means of linguistic co-occurrence networks?
- **Idea**: Replace the standard text mining feature sets with linguistic network measures for the purpose of text classification
  - Reduce huge feature-space

Dataset

- **150** equal-sized Croatian texts divided in two classes:
  - 75 literature texts
  - 75 blog texts

- Linguistic distinction between literature and blog
  - Literature texts: segments from 7 different books written in or translated to Croatian language
  - Blog texts: collected from two popular Croatian blogs

- $\sim$ 10000 words per text
- Lemmatized
- Stopwords removed

Network Construction

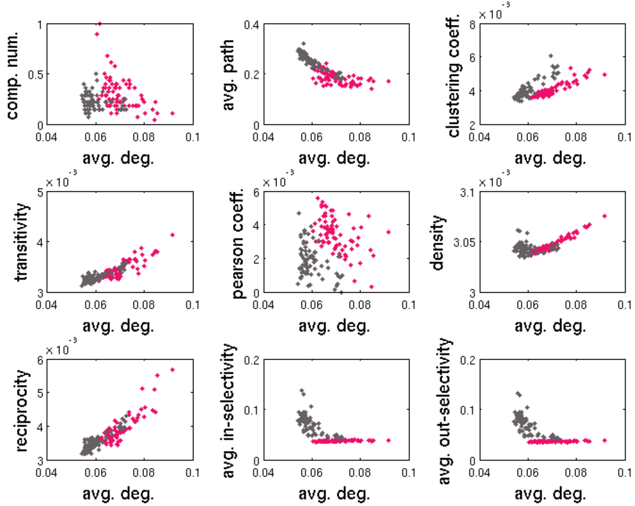
- **150** different co-occurrence **networks**
- One network for each text in the dataset
  - all weighted and directed
- The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a text

Feature set

- For each network we computed a set of 15 measures
- Correlated features (> 0.8) were removed
  - diametar, radius, mean in- and out- degree, avg. node conn.
- **10 measures left**:
  - number of components,
  - average degree,
  - average path length,
  - clustering coefficient,
  - transitivity,
  - degree assortativity,
  - density,
  - reciprocity,
  - average in-selectivity,
  - average out-selectivity

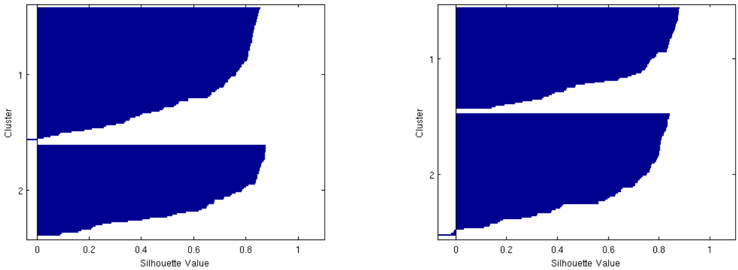
- All features are rescaled to [0 - 1]

Raw data visualisation



Experiments: k-means clustering

- Cosine & correlation point-to-centroid distances



Mean silhouette values:

k/distance	cosine	correlation
2	**0.6550**	**0.6486**
3	0.6494	0.6472
4	0.5659	0.5726

Experiments: Classification

- Train set size: **135**
- Test set size: **15**
- 10-fold cross-validation

- Classification methods:
  - Support vector machines,
  - Classification trees,
  - Naive Bayes,
  - k-nearest neighbors,
  - Linear discriminant analysis (+QDA)

Experiments: Classification II

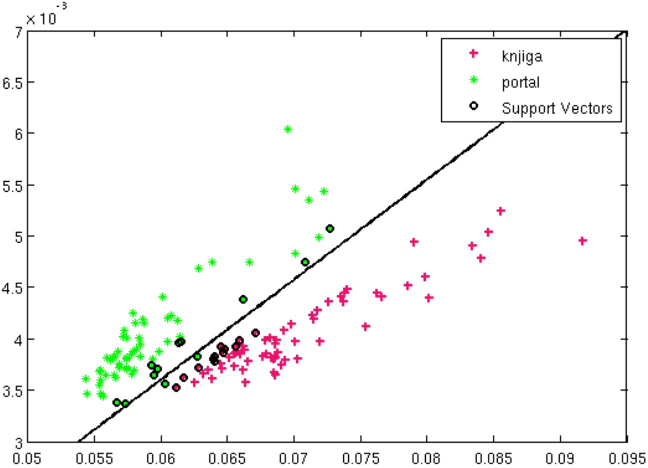
- There are tradeoffs between several characteristics of classification algorithms

Algorithm	Predictive Accuracy	Fitting Speed	Prediction Speed	Memory Usage
Trees	Medium	Fast	Fast	Low
SVM	High	Medium	Medium	Medium
Naive Bayes	Medium	Medium	Medium	Medium
Nearest Neighbor	Medium	Fast	Medium	High
Discriminant Analysis	High	Fast	Fast	Low

http://www.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html

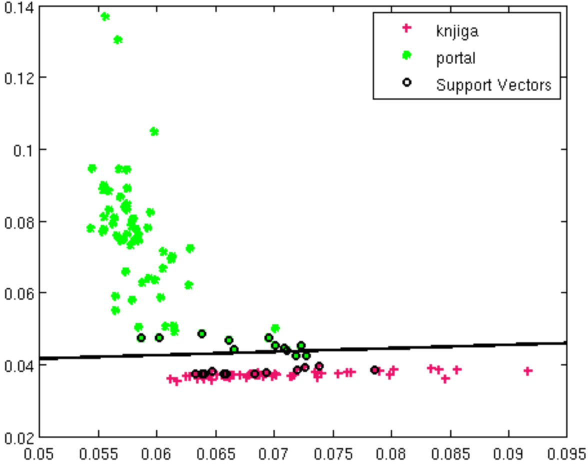
Support vector machines

Cross-validation error: **0.0067%** on all features



Avg. degree - clustering coeff. plot

Support vector machines II



Avg. degree - avg. in-selectivity plot

Naive Bayes

Naive Bayes

- Misclassification error: **0** (100% success rate)
- Cross-validation error: **0.0067**
- Confusion matrix:
  7 0
  0 8

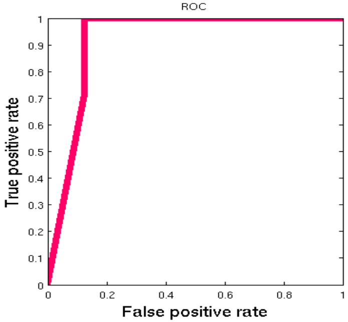
- The same error as for SVM classification

Nearest Neighbors

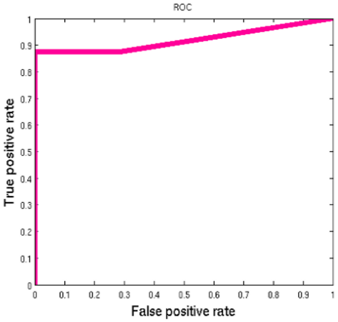
Receiver Operating Characteristic curve

Positive class:

literature



blog



AUC: **0.9196**

Misclassification error: **0.0067**

CV error: **0.0533**

LDA (+QDA)

LDA

- Misclassification error: **0** (100% success rate)
- Cross-validation error: **0** (100% success rate)
- Best result!

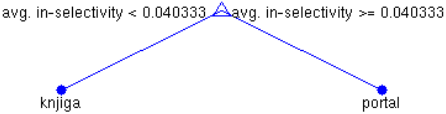
QDA

- Misclassification error: **0.0067**
- Cross-validation error: **0.0133**
- Confusion matrix:
  6 1
  0 8

Classification Tree

Decision tree for classification:

if avg. in-selectivity < 0.0403
then *literature*
elseif avg. in-selectivity $>= 0.0403$
then *blog*
else *literature*

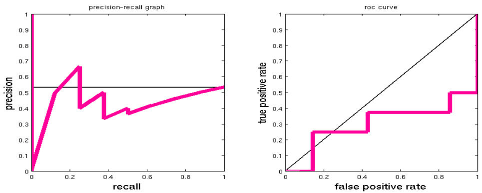


S. Šišović, S. Martinčić-Ipšić, and A. Meštrović. "Comparison of the language networks from literature and blogs." Mipro 2014.

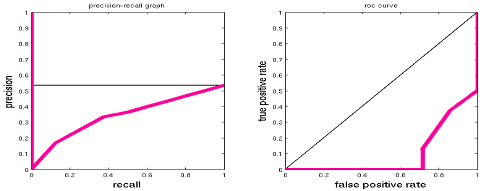
arXiv:1405.2702 (2014)

Precision-Recall & ROC

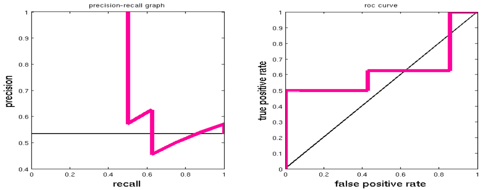
Average degree



Number of components



Clustering coefficient

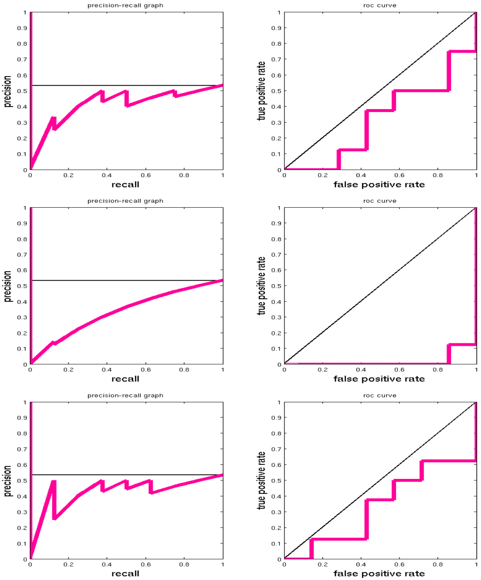


Precision-Recall & ROC

Transitivity

Degree assortativity

Density

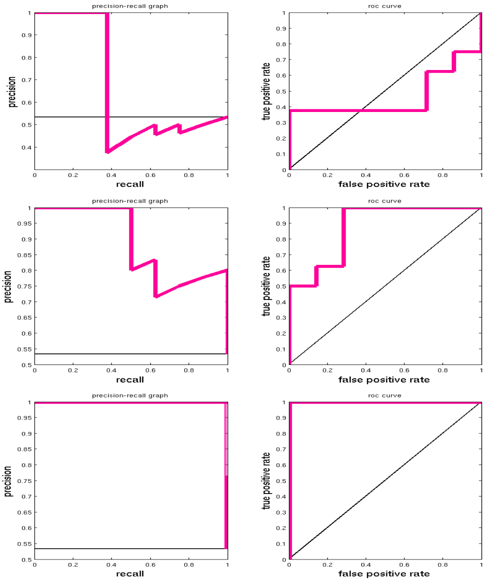


Precision-Recall & ROC

Reciprocity

Avg. path length

Avg. in- and out-selectivity



Conclusion

- We replaced the standard text-mining features with **complex network measures**
  - 10 features (15 initial)
  - reduced traditional NLP feature set by ∼ 10000:10
- Preliminary experiment over-simplified: only two text types
- Classification methods: SVM, classification trees, Naive Bayes, Knn, LDA, QDA
  - all classifiers showed similar results for simple net-based classification
  - misclassification errors **less then 1%**
- **Average selectivity measure** - most useful feature for predicting the correct text type and reducing the misclassification rate
  - precision, recall and ROC indicate that the average node selectivity has potential **to capture the structural differences** between two classes of texts

Future work

- Include & explore additional complex network measures
- Test on different text collections, different text sizes
- More classes, more complex problems
- Develop the method for network-based topic classification, fine-grained text type classification, text genre differentiation, etc.
- Toward possible text quality evaluation?

Toward a Complex Networks Approach on Text Type Classification

Suggestions, Remarks, Questions most welcome

{dmargan,amestrovic,marinai,smarti}@uniri.hr