UNIVERSITY OF TARTU
Institute of Computer Science

Extracting facts from medical texts

Robert Roosalu, supervisor Sven Laur.
University of Tartu, Faculty of Science and Technology, Institute of Computer Science, MSc Computer Science

STACC

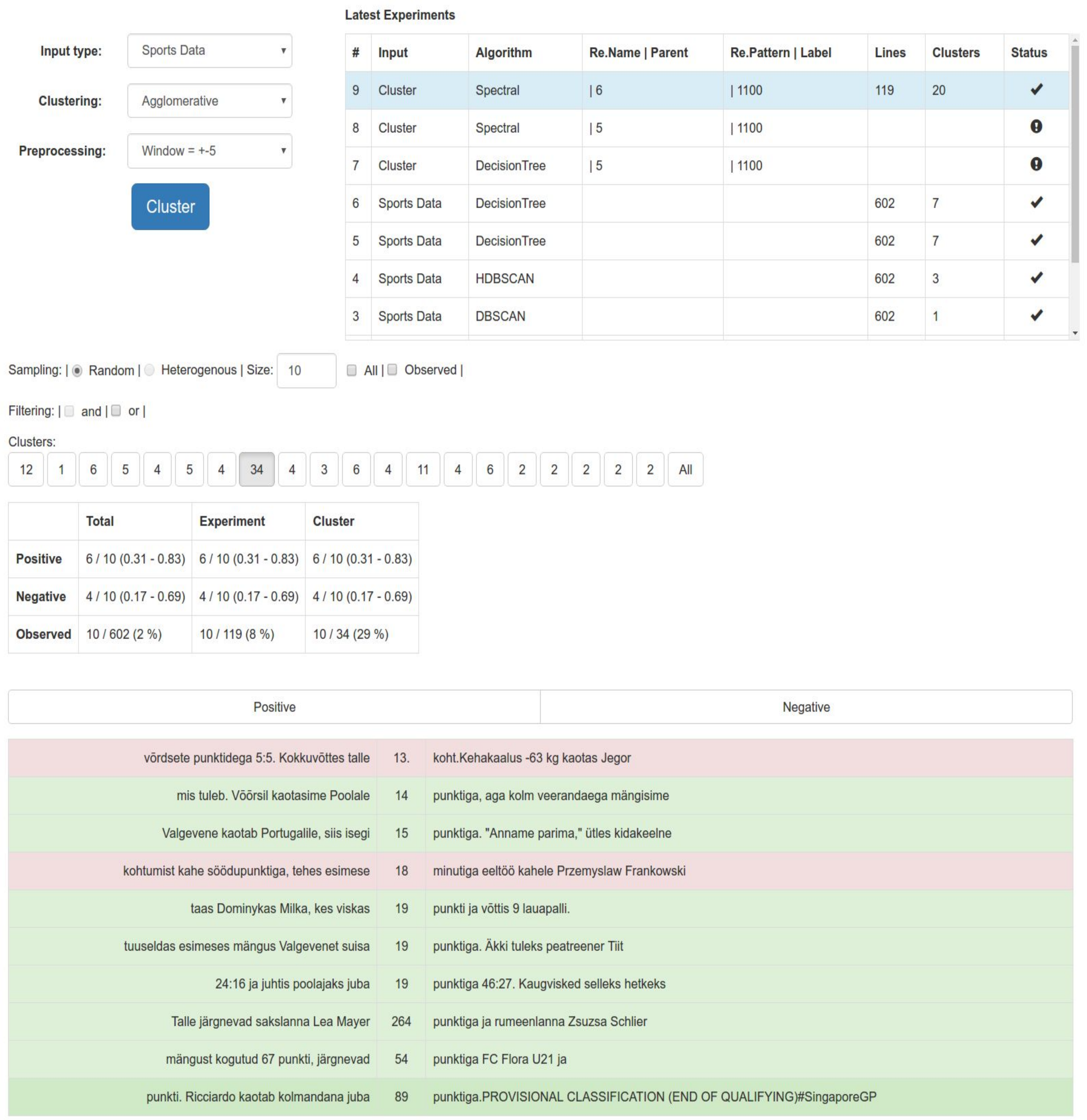
Study IT in .ee
sponsored by Skype

Introduction

Estonian *Geenivaramu* has built a large corpora of free-form texts from doctors, describing patient's diagnosis, treatment, etc. Successfully mining factual knowledge from this kind of data would lead us closer towards advancements in personal medicine, automated clinical trials, etc.

One chosen approach is via human defined patterns, such as regular expressions, or context free grammars. The results of the patterns are too numerous to be manually evaluated. To enable such evaluation, we have built PatternExaminer. The tool clusters the contexts of the extracted facts, which are then sampled for the final evaluation.

For the success of PatternExaminer, we need the highest quality text clustering capabilities.



Input type: Sports Data
Clustering: Agglomerative
Preprocessing: Window = +-5
Cluster

Latest Experiments

#	Input	Algorithm	Re.Name | Parent	Re.Pattern | Label	Lines	Clusters	Status
9	Cluster	Spectral	| 6	| 1100	119	20	✓
8	Cluster	Spectral	| 5	| 1100			❗
7	Cluster	DecisionTree	| 5	| 1100			❗
6	Sports Data	DecisionTree			602	7	✓
5	Sports Data	DecisionTree			602	7	✓
4	Sports Data	HDBSCAN			602	3	✓
3	Sports Data	DBSCAN			602	1	✓

Sampling: | ● Random | ○ Heterogenous | Size: 10 □ All | □ Observed |

Filtering: | □ and | □ or |

Clusters:
12 1 6 5 4 5 4 34 4 3 6 4 11 4 6 2 2 2 2 2 All

	Total	Experiment	Cluster
Positive	6 / 10 (0.31 - 0.83)	6 / 10 (0.31 - 0.83)	6 / 10 (0.31 - 0.83)
Negative	4 / 10 (0.17 - 0.69)	4 / 10 (0.17 - 0.69)	4 / 10 (0.17 - 0.69)
Observed	10 / 602 (2 %)	10 / 119 (8 %)	10 / 34 (29 %)

Positive | Negative

võrdsete punktidega 5:5. Kokkuvõttes talle	13.	koht.Kehakaalus -63 kg kaotas Jegor
mis tuleb. Võõrsil kaotasime Poolale	14	punktiga, aga kolm veerandaega mängisime
Valgevene kaotab Portugalile, siis isegi	15	punktiga. "Anname parima," ütles kidakeelne
kohtumist kahe söödupunktiga, tehes esimese	18	minutiga eeltöö kahele Przemyslaw Frankowski
taas Dominykas Milka, kes viskas	19	punkti ja võttis 9 lauapalli.
tuuseldas esimeses mängus Valgevenet suisa	19	punktiga. Äkki tuleks peatreener Tiit
24:16 ja juhtis poolajaks juba	19	punktiga 46:27. Kaugvisked selleks hetkeks
Talle järgnevad sakslanna Lea Mayer	264	punktiga ja rumeenlanna Zsuzsa Schlier
mängust kogutud 67 punkti, järgnevad	54	punktiga FC Flora U21 ja
punkti. Ricciardo kaotab kolmandana juba	89	punktiga.PROVISIONAL CLASSIFICATION (END OF QUALIFYING)#SingaporeGP

Figure 2. User interface of PatternExaminer.

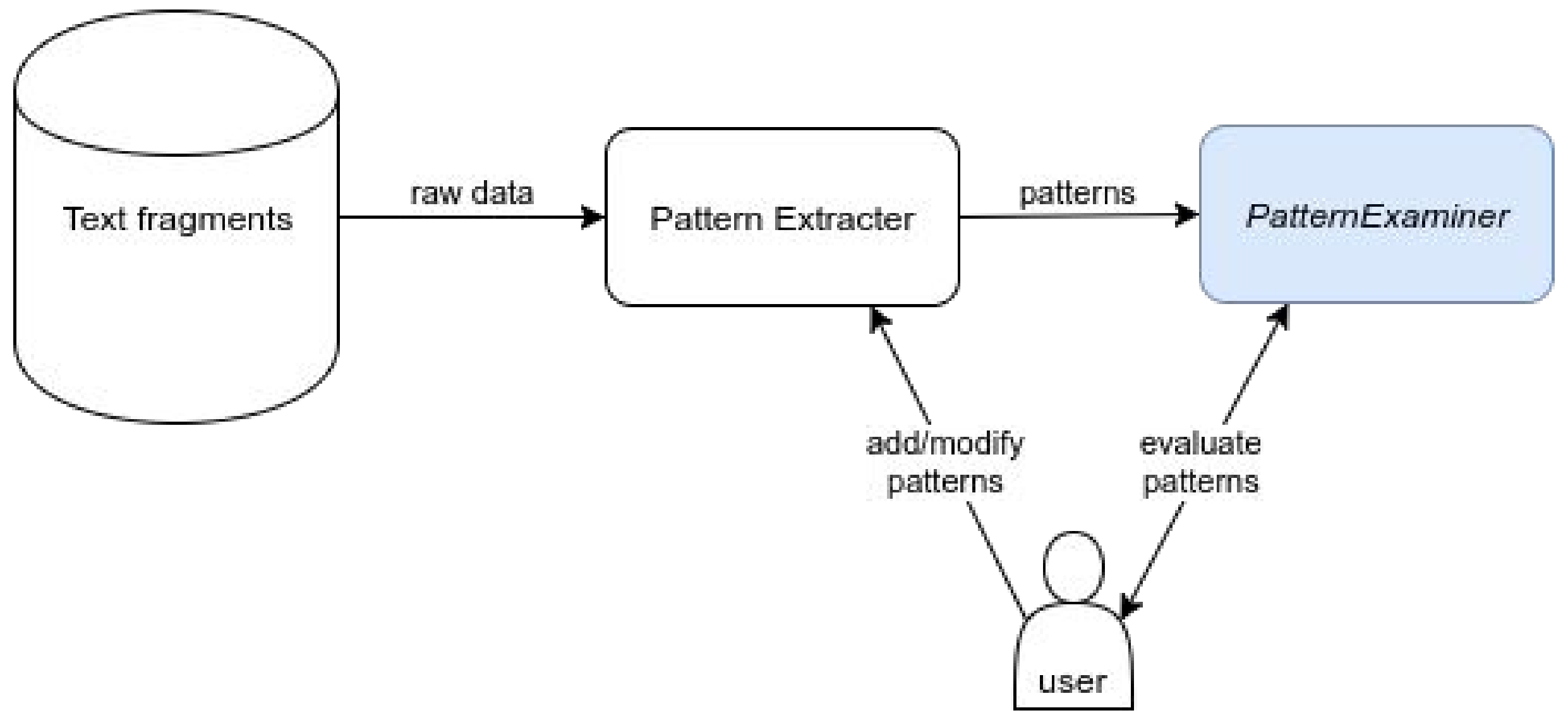


Figure 1. Workflow for validating the defined patterns.

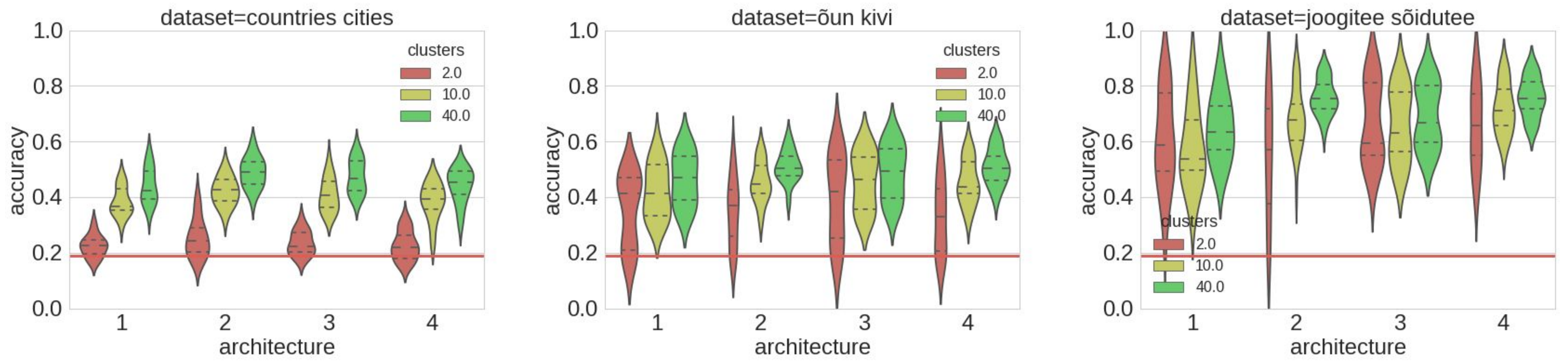


Figure 3. Results of recurrent autoencoder experiments, Spectral clustering.

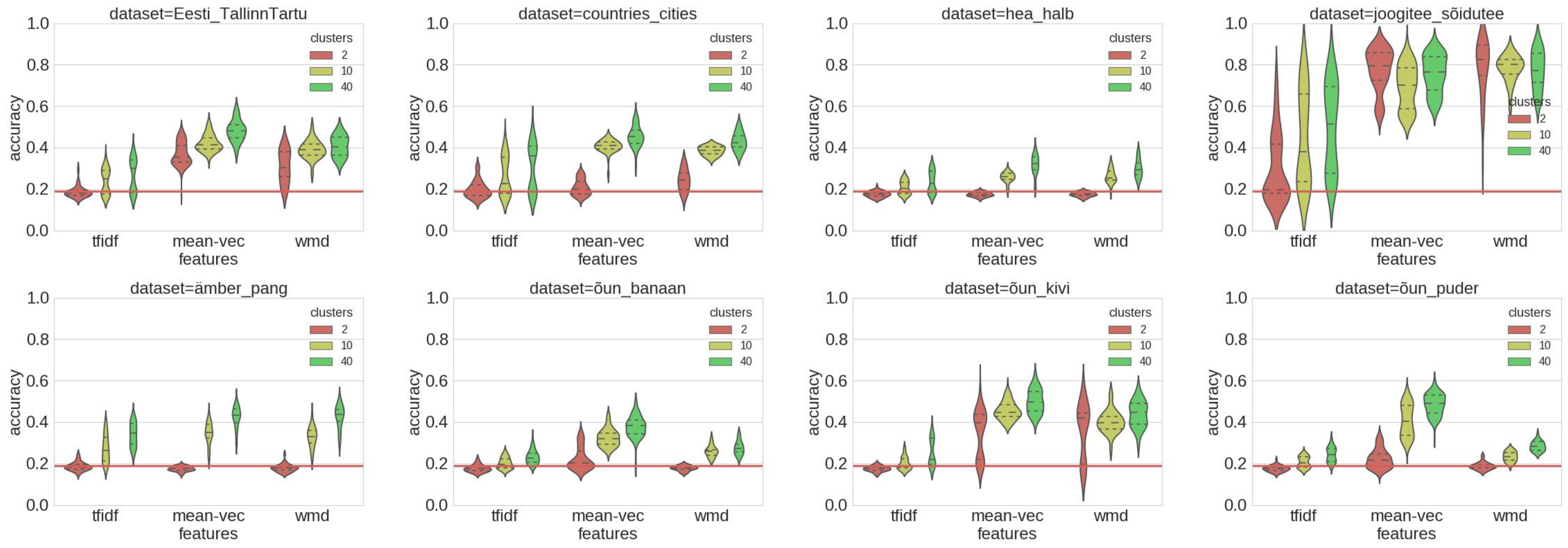


Figure 4. Various results, Spectral clustering.

Experiments

For the dataset, eight different pairs of words were extracted from the Estonian Reference Corpus. Grid search was used over an array of various methods, parameters and hyperparameters, leading to a total of around 20000 different clustering experiments.

For evaluating the quality of clusterings, a custom set of metrics emulating the usage of PatternExaminer was employed.

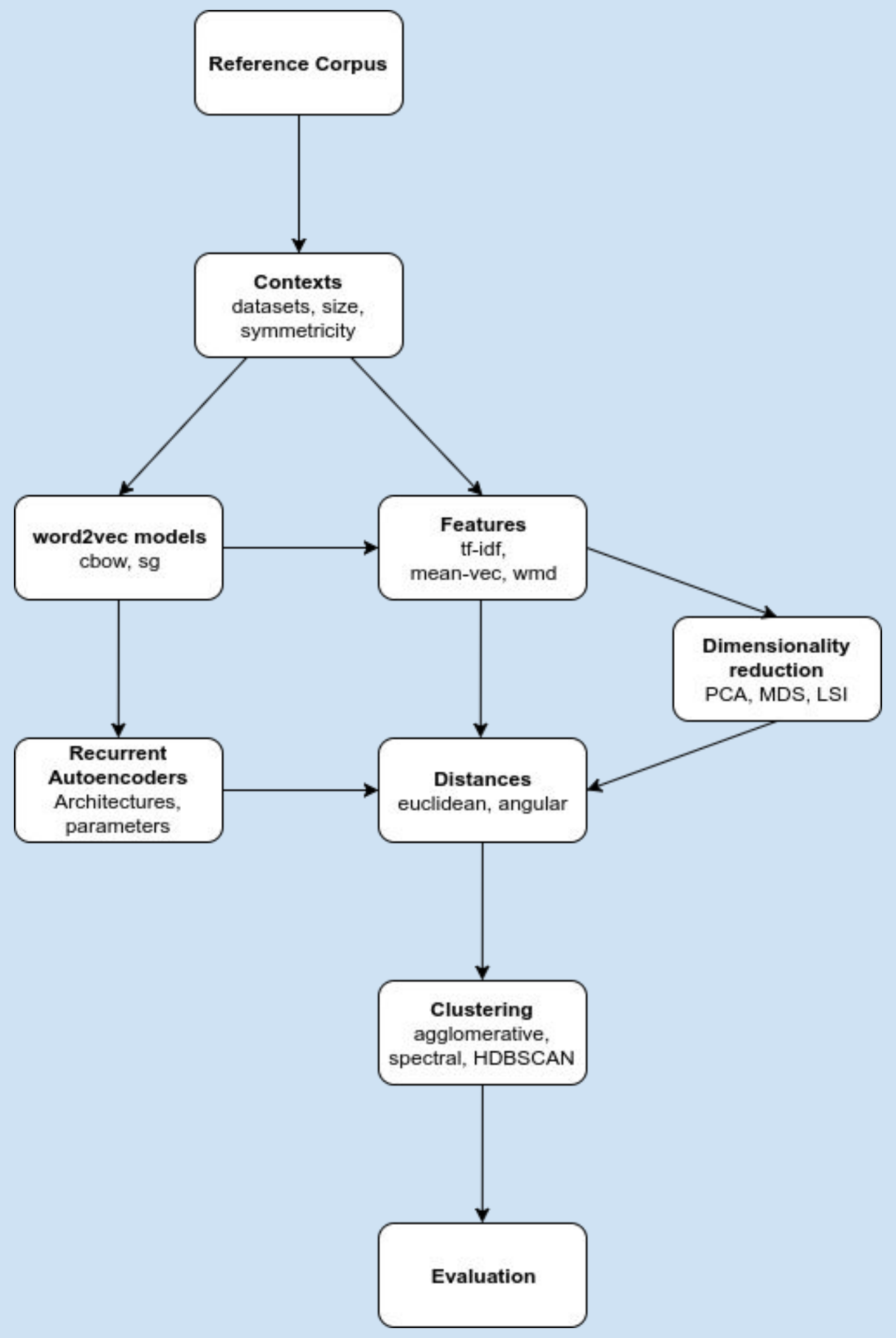


Figure 5. Experiment pipeline.

Results

For analyzing such a large amount of experiment results, we used categorical faceted violin plots.

It was found, that spectral clustering is best for clustering textual data, in comparison with Agglomerative and HDBSCAN.

For the classical methods, a good context window is symmetric, but the size doesn't matter. Skip-gram word embeddings, averaged to the mean vectors and using angular distance yields the most separable embedding space. Dimensionality reduction did not prove useful in the experiments.

Simple recurrent neural network architectures achieved on-par or better results, than the mean-vector approach.

Future work

Building a test set on the *Geenivaramu* dataset, to evaluate all following models.
Transfering the algorithms onto the *Geenivaramu* dataset, verifying if the current insights hold.
Employing state-of-the art RNN architectures for the embedding.

Master's thesis *Context Embeddings for Natural Language Clustering*
https://github.com/andmekaevur/thesis-experiments
https://github.com/estnltk/pattern-examiner