A Semi-Supervised Learning System for Micro-Text Classification

Thomas Ott*, Markus Christen†, Thomas Niederberger*, Reto Aebersold‡,Suleiman Aryobsei§, Reto Hofstetter§
* ZHAW Zurich University of Applied Sciences, Switzerland, Email: ottt@zhaw.ch
†University of Zurich, Switzerland
‡ Atizo AG, Bern, Switzerland
§University of St. Gallen, Switzerland

***Abstract*—Modern communication platforms are sources of large samples of micro-texts that are in need of machine text processing for text classification and interpretation. On the Swiss innovation platform *Atizo*, contributors submit ideas or solutions to problems posed by companies. In such a crowdsourcing process, hundreds of people contribute a large amount of micro-text data that needs to be structured already during the process of idea generation in order to avoid repetitions and to optimize the solution space. Due to low word-count and unstructured writing, micro-texts pose a challenge for automatized text processing. Technically, the goal is to partition a growing set of micro-texts $\mathcal{T} = \{t_1, ..., t_n\}$ into m topical groups $G_j \subset \mathcal{T}, j \in \{1, ..., m\}, (\cup_j\ G_j = \mathcal{T}, G_k \cap G_l = \emptyset)$, where each text is characterised by a set of words $t_i = \{w_{i_1}, ...w_{i_{n_i}}\}$ and the corresponding word counts (bag-of-words model, e.g. [1]) and each group G_j is associated with a set of terms or key words k_{G_j} defining the topic of the group, i.e.**

$$t_i \to G_j,\ k_{G_j} \subset\ \cup_\alpha\, t_\alpha \quad \textbf{with} \quad t_\alpha \in G_j$$

according to some suitable optimality criterion.

In our research, we addressed the whole chain of issues related to data pre-processing (where text enrichment plays a major role), data classification, data visualization, as well as solution benchmarking and process improvement. Our research eventually culminated in the development of a complete system for machine-supported clustering and classification of micro-texts on the innovation platform *Atizo*. The system combines a bottom-up clustering approach with a top-down classification approach (Fig 1). This leads to a kind of semi-supervised learning procedure, where an unsupervised learning step to identify the best cluster(s) is combined with a supervised control step to define a classifier. For the clustering step, we implemented a PCA-based approach (Hebbian clustering [2]) that yields clusters and a corresponding characterization by means of key words. In the supervised step, a human supervisor intervenes by selecting the most significant key words for the best cluster (as assessed by a suitable criterion). These key words are then used to generate a classifier.

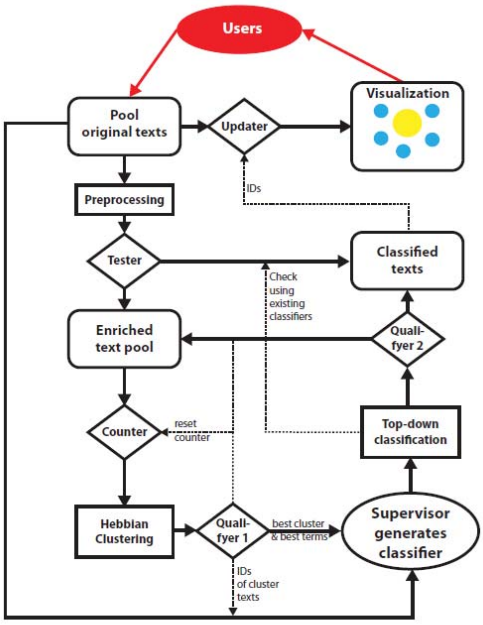

Fig. 1. Flowchart of the entire classification system

The entire procedure of clustering and classification is run through again as long as new data is entering the system or significant clusters can be found.

In our contribution, we present the technical details of our system and discuss the results from preliminary studies on the innovation platform *Atizo*.

Acknowledgment: This project has been supported by KTI grant 12747.1 PFES-ES.

REFERENCES

[1] M. Steyvers, T. Griffiths. Probabilistic Topic Models. In *Latent Semantic Analysis: A Road to Meaning*. (Lawrence Erlbaum), 2007.

[2] T. Niederberger, N. Stoop, M. Christen, T. Ott. Hebbian Principal Component Clustering for Information Retrieval on a Crowdsourcing Platform. In *Proceedings of NDES'12*, 2012.