NOLTA, IEICE

Paper

Micro-text classification between small and big data

Markus Christen [1 a)], *Thomas Niederberger* [2], *Thomas Ott* [2], *Suleiman Aryobsei* [3], *and Reto Hofstetter* [4]

[1] *University Research Priority Program Ethics, University of Zurich Zollikerstrasse 117, 8008 Zurich, Switzerland*

[2] *Center for Predictive & Bio-Inspired Modeling, Zurich University of Applied Sciences, Einsiedlerstrasse 31a, 8820 Wädenswil, Switzerland*

[3] *Center for Customer Insight, University of St. Gallen Bahnhofstrasse 8, 9000 St. Gallen, Switzerland*

[4] *Institute of Marketing and Communication Management, University of Lugano, Via G. Buffi 13, CH-6904 Lugano, Switzerland*

[a)] *christen@ethik.uzh.ch*

Received January 27, 2015; Revised June 25, 2015; Published October 1, 2015

Abstract: Micro-texts emerging from social media platforms have become an important source for research. Automatized classification and interpretation of such micro-texts is challenging. The problem is exaggerated if the number of texts is at a medium level, making it too small for effective machine learning, but too big to be efficiently analyzed solely by humans. We present a semi-supervised learning system for micro-text classification that combines machine learning techniques with the unmatched human ability for making demanding, i.e. nonlinear decisions based on sparse data. We compare our system with human performance and a predefined optimal classifier using a validated benchmark data-set.

Key Words: micro-text, text enrichment, text classification, innovation management, social media platforms, semi-supervised learning

1. Introduction

The pervasive use of internet services such as *Facebook*, *Twitter*, *Yahoo! Answers*, crowdsourcing platforms, or micro-blogging services allow for an unprecedented access to user generated content that can be used for various social data mining applications [1]. Many online platforms encourage users to submit micro-texts, such as user-generated content on social media portals [2]. By the term "micro-text" we refer to texts that have a rudimentary grammatical structure but are mostly unstructured and consist of one or only a few sentences. Paradigmatic examples of micro-texts are *Twitter* tweeds, texts emerging from short message services, or comments on news portals.

Analyzing these micro-texts has a great scientific potential in psychology [3] and social sciences [4], but is methodologically challenging [5] and typically heavily depends on machine text processing

Nonlinear Theory and Its Applications, IEICE, vol. 6, no. 4, pp. 1–14 DOI: 10.1588/nolta.6.1

to cope with the large amounts of published texts. A fundamental step in social data mining is text classification, which may involve both identifying texts that match predefined classes (top-down classification) as well as finding an unknown classification scheme that fits the data (bottom-up classification). This endeavor relies on a rich tradition of text classification in various fields, e.g., library science or bibliographic classification theory [6], where the concept of faceted classification – the discovery of unknown classes in a set of texts – already emerged in the pre-digital times. The human ability to grasp the semantics of texts allows for sophisticated and relatively fast classifications as long as the number of texts is low, e.g., new entries in a public library.

The nature of micro-texts adds substantial complication to text classification due to the limited length of those texts, the violation of grammatical or stylistic conventions, the low text structure by traditional natural language processing definitions – usually, only a source attribution (author) and a time stamp are available –, and the pervasive abbreviations and coined acronyms [7, 8]. Nevertheless, various applications like spam detection on *Twitter* [9] or digital disease surveillance [10] have been developed in recent years. The approaches presented in this prior work are promising, but typically depends on high numbers of texts to allow efficient machine learning. For instance, approximately 0.5 million Tweeds were used to assess vaccination sentiments within online social media [11]. The same holds for the application of popular classification methods like Latent Dirichlet Allocation [12] to micro-texts classification. For example, Ramage, Dumais and Liebling [13] used more than eight million *Twitter* Tweeds to train their model.

A major reason for the successful application of machine learning techniques for large scale text classification problems is the fact that text classification is often a nearly linearly separable problem [14]. However, if the number of training texts is small and in the face of very high-dimensional text spaces, the decision boundaries are under-determined which leads to a bad generalization performance of classifiers. Humans, in contrast, can account for higher-order term interactions upon their ability to grasp the semantics of the texts. Considering higher-order interactions introduces a nonlinear capacity [14] that serves as a base for reliable classifications.

Thus, the problem of micro-text classification – and text classification in general – can be conceptualized between two poles (Fig. 1): If the number of texts is small (a few dozen), humans typically outperform machine classification because there is not sufficient text data available to allow for statistical machine learning and because humans have a quick, intuitive understanding of the semantics of micro-texts despite their unstructured form and do not need to rely their decision on the whole text corpus [15]. If the number of texts is large, however, humans are overburdened, whereas machine learning usually improves with increasing amount of text data, although the improvement saturates after a certain number, depending on the method used [16]. The problem is how to deal with medium-sized sets of micro-texts that are in the order of a few hundred to a few thousand, which are too large for humans to be overseen in a reasonable amount of time but that do not offer enough data for

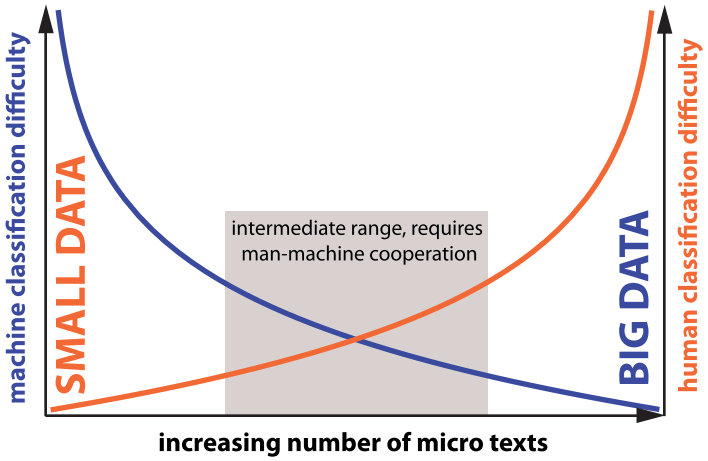

Fig. 1. Conceptualizing the problem of classifying medium-sized sets of micro-texts.

efficient machine learning. An example would be to classify comments posted on media websites for popular articles, typically a few hundred micro-texts. Our contribution provides answers to this question by taking into account both fundamental types of classification – bottom-up and top-down.

Finding solutions for medium-sized text sets also have economic implications, as human evaluation and classification of ideas usually comes at a substantial cost. For the evaluation of ideas, usually multiple experts are hired who then evaluate multiple ideas on different dimensions such as their novelty and effectiveness [17]. Girotra et al. [18] for instance hired 41 MBA students who each evaluated between 206 and 237 different ideas. Hofstetter et al. [19] analyzed 601 ideas generated during an ideation experiment. Each idea was evaluated by 13 experts who spent 1 minute on average per evaluation resulting in a total workload of 130 hours. Classifying ideas is even more time consuming as not only each idea but also each pair of ideas has to be assessed. Kornish and Ulrich [20] provide an example of this workload. A set of 290 ideas requires $(290289/2) = 41{,}905$ human similarity judgments. With three raters for each pair of ideas and if each judgment took only 15 seconds, the human approach would require 175 hours of rater effort, more than a month of work, which would be prohibitively time consuming and costly.

There are two options to deal with medium-sized sets of micro-texts. One option is to employ text preprocessing and text enrichment such that the information content of a single text is better accessible for machine classification. Standard preprocessing procedures include stop word elimination, orthographic error replacement, stemming, lemmatizing (setting a word to its base form), and splitting of compound words. Latter is of particular important in German texts due to the common use of compound words; [21]. Enrichment techniques refer to the use of synonym databases [22] or translations of the text [23]. A second option refers to the combination of human and machine intelligence such that the strengths of each of them can be used in an optimal way.

To test both the use and performance of text preprocessing and its combination with human intelligence, we use medium sized sets of micro-texts from the online ideation platform *Atizo.com*. On *Atizo.com*, companies seeking new ideas ("seekers") host ideation contests for their innovation problems and users ("solvers") suggest solutions to these problems by submitting their ideas in the form of micro-texts. The ideas typically span a wide solution space that maximizes the potential to find the best solution for the problem posed. Seekers reward the best solvers for their ideas as an incentive to participate. As an ideation process may yield hundreds of contributions, both the seeker and the solver face the tasks of structuring, classifying and evaluating the contributions. Analysis is not only complicated by the short length of the micro-texts, but also by a large number of submissions. A typical problem is that humans lose the overview on the submitted ideas as soon as the number of ideas exceeds a certain amount; typically more than 50. As a consequence, the repetition of ideas hinders the optimal exploration of the solution space. To solve this problem, a classification of ideas is required that outlines the structure of the idea space. This is an example of a bottom-up classification that is challenging both for humans due to efficiency, as a real-time classification should be achieved because the text set grows in time and for machines due to the insufficient number of texts for machine classification.

In what follows, we first outline the conceptual background of micro-text classification. Second, we describe in detail the various steps of our enrichment and classification procedure. Third, we present performance results based on real data emerging from the ideation platform *Atizo.com*. Finally, we discuss the relevance of our findings for the general problem of classification on medium-sized micro-text sets.

2. Conceptual background and hypothesis

Technically, the problem of micro-text classification under consideration can be stated as follows: Given is a set of micro-texts $M(T)$ that grows over time by adding new texts in discrete variable time steps

$$M(T) = \bigcup_{t=1}^{T} \{m(t)\} \tag{1}$$

where T is the current total number of micro-texts and $m(t)$ is the text that has been added at time step t. Each text is characterized by a set of words $m(t) = \{w_{t_1}, ..., w_{t_n}\}$ which, after preprocessing (comprising text enrichment, see below), may include more words than the original document. The goal is to partition $M(T)$ into k topical groups

$$G_j(T) \subset M(T), j \in \{1, \ldots, k\}, G_k(T) \cap G_l(T) = \emptyset \tag{2}$$

in such a way that each group is associated with a limited set of terms or key words $k_{G_j(T)} \subset \bigcup_t m(t)$ with $m(t) \in G_j(T)$ and these key words alone allow for a best possible classification or assignment

$$m(t) \rightarrow G_j(T) \text{ if } m(t) \in G_j(T) \tag{3}$$

According to this formulation, the problem inherently comprises the two tasks of (1) a clustering to identify relevant topical groups $G_j(T)$ (bottom-up) and of (2) a key word-based classification $m(t) \rightarrow G_j(T)$ (top-down) to assess and use the clustering results. The top-down and bottom-up approaches have to be understood as complementary parts that, to some degree, are dependent on each other.

We hypothesize that both approaches need to be integrated in a final solution. In addition, text preprocessing and enrichment are likely elements of the procedure due to the low number of micro-texts. Finally, human interventions may be useful to improve the results of micro-text classification significantly, but should be as small as possible to allow for a high efficiency of the system. We test these possibilities – enrichment, bottom-up classification, top-down classification – independently as well as their integration in a complete system that emerges as a result of our investigations.

3. Procedure

3.1 Data set description and benchmark creation

For our experiments, we used micro-texts of three ideation contests from *Atizo.com*. In the first contest, participants were asked to provide ideas for how a museum could become more attractive and would attract more visitors (394 texts by 154 solvers). In the second contest, participants were asked to provide ideas for how a convenience food producer could increase its range of products (314 texts by 129 solvers). In the third contest, participants were asked to provide ideas for how a new alcoholic beverage could be advertised (396 texts by 117 solvers). The texts were mostly in German, some were in English or French. Table I provides three examples of original texts as well as the effect of a selected type of preprocessing (stop word elimination, word splitting and lemmatizing, see Table III for details).

Each data set was hand-clustered by one author (M.C.) to identify clear groups of ideas within each project. Among those groups, 100 German micro-texts per project were selected by all authors such that for each project, the majority of the ideas belong to four clearly distinguishable groups of different sizes, while a minority of ideas belong to neither group ("noise texts"). In this way three test-sets were generated as baseline classification. Each idea text consisted of a title, some key-words chosen by the solvers and the text describing the idea. Often, titles and keywords were bad descriptors of the actual content of the idea. Thus, privileging these parts of the text was of no use and we represented all ideas as single word bags; i.e., the set of all words that are contained in the text including repetitions. The mean numbers of words per bag were 58 for project 1, 56 for project 2 and 61 for project 3.

For validating the baseline classification, each group for each project has been described explicitly by a short text such that a clear instruction emerged how to attribute each single text to one group. For each project, 10 subjects classified all 100 ideas along these instructions. At least 8 out of 10 subjects had to agree that a single idea belongs to a specific group. In this way, a benchmark classification of the 100 ideas of each project was achieved. Table II outlines the size of each group as well as the number of "noise ideas", i.e., ideas unrelated to any group.

This prior classification shown in Table II served as the benchmark clustering C_{ref}. We evaluated the quality of a clustering result using the Jaccard coefficient J based on a clustering result C and the benchmark C_{ref} as follows

Table I. Example texts. One typical example of each ideation contest (the English translation approximates the sloppy writing style of the German original texts). Title are in bold, keywords are in italic; both title and keywords have been defined by the solvers.

Contest	English translation	German original text	Text after type 6 preprocessing
Contest 1, text of the group "idea refers to children"	**Children's Birthday at Museum** *children, museum, birthday* Offer children's birthday parties at museums. Depending on the age, the party could include an art quiz (rally through the museum), workshops (be creative), a party, or visit the museum at night.	**Kindergeburtstag im Museum** *kinder, museum, geburtstag* Speziell für verschiedene Altersgruppen werden Kindergeburtstage im Museum angeboten. Kunstquiz (Rally durchs Museum), Workshop (selber gestalten), Party im Museum, Museum bei Nacht	kind geburtstag museum kind museum geburtstag altersgruppe kind geburtstag museum angebot kunst quiz rally museum workshop selber gestalt party museum museum nacht
Contest 2, text of the group "idea refers to soups"	**Soup-machine** *soup, machine, coffee machine, at home, office, cube, capsule* A soup-machine for home or office. Like a coffee machine, but for soups. The soup could be filled into cubes or capsules.	**Suppen-Automat** *suppen, automat, kaffeemaschine, zuhause, büro, würfel, kapsel* Ein Suppenautomat für zuhause oder im Büro. Wie eine Kaffeemaschine, nur für Suppen. Die Suppen könnten in Würfel- oder Kapselform eingefüllt werden.	suppe automat suppe automat kaffee maschine zuhause büro würfel kapsel suppe automat zuhause büro kaffee maschine suppe suppe würfel kapsel form einfüllen
Contest 3: text of the group "idea refers to bottle shape"	**Special bottle shapes** *bottle shape, form, toast* Special bottle shapes for couples, two bottles fit to each other like yin and yang, or fit together forming a star shape for groups. Makes toasting a special highlight.	**Spezielle Flaschenformen** *flaschenform, form, anstossen* Spezielle Flaschenformen für Pärchen, bei denen genau zwei Flaschen (wie bei Yin und Yang) ineinanderpassen oder in Sternform fär Gruppen. Macht das Anstossen zum besonderen Highlight.	speziell flasche form flasche form form anstossen speziell flasche form pärchen flasche yin yang ineinander passen sternform gruppe machen anstossen highlight

Table II. Benchmark classification. Distribution of texts of three ideation contests from *Atizo.com* in four classes and noise texts.

	Size of group 1	Size of group 2	Size of group 3	Size of group 4	# of noise texts
Project 1	10	9	21	22	38
Project 2	9	10	17	20	44
Project 3	9	10	15	24	42

$$J(C, C_{ref}) = \frac{a}{a+b+c} \tag{4}$$

where a is the number of pairs of items that are both in C and C_{ref} in same clusters, b is the number of pairs that are only in C in same clusters, and c is the number of pairs that are only in C_{ref} in same clusters. $J(C, C_{ref})$ ranges between 0 and 1, with a value close to 1 signifying a similar clustering. When calculating the Jaccard coefficient, we removed all noise texts from C in order to assess to what extend a specific method was able to reproduce the initial clustering. The reasoning for this was that the noise texts are not a group with a specified meaning. The only unifying characteristic of noise texts was that they did not belong to any of the predefined groups. Detecting potential sub-structure in the noise texts would be dependent on the method used and would make it impossible to compare all results unambiguously.

3.2 Text preprocessing and enrichment

The first goal of the study was to evaluate the effect of preprocessing and text enrichment of the clustering result. We measured the improvement in terms of the Jaccard coefficient compared to the unprocessed texts after elimination of standard stop words like articles "der, "die", "das" etc. (= set 1). Table III provides an overview of the preprocessing and enrichment techniques used in the study.

For the enrichment technique *Translation*, the texts have been translated automatically using *Google translate* and standard stop words have been removed. *Lemmatizing* means that all words have been

Table III. Overview of text processing procedures. *Abbreviations: S: Stemming, T: Translation, H: word splitting, L: lemmatizing, Y: synonym enrichment. The sources mentioned in the column "Description" provide detailed information on the functioning of each tool.

Type	Abbreviation*	Description
1	-	Raw texts after elimination of standard stop words
2	S	Stemming of the texts of type 1 (tool: German Porter Stemmer; [24])
3	T	Translation of the texts of type 1 in English and French (tool: *Google translate*)
4	H	Word splitting of the texts of type 1 (tool: *jWordSplitter*; available at: `http://www.danielnaber.de/jwordsplitter/`)
5	H & T	Translation of the texts of type 4
6	H & L	Lemmatizing of the texts of type 4 (tool: German morphology lexicon; available at: `http://www.danielnaber.de/morphologie`)
7	H & L & T	Translation of the texts of type 6 (English & French)
8	H & L & Y	Enrichments with synonyms of the texts of type 6 (tool: German *OpenThesaurus*; available at `http://www.openthesaurus.de/`)
9	H & L & Y & T	Translation of the texts of type 8 (English & French)
10	H & S	Stemming of the texts of type 4
11	H & L & S	Stemming of the texts of type 6
12	H & L & Y & S	Stemming of the texts of type 8

replaced by the base form as nouns, adjectives or verbs. For enrichment with *Synonyms*, we added for each word of a word bag, for which synonyms were available on the German *OpenThesaurus* database (`www.openthesaurus.de`), all synonyms to the word bag. After applying each of these steps, we performed bottom-up clustering as described in the next section and top-down classification as described in Section 3.4 to identify clusters.

3.3 Bottom-up hebbian clustering

For clustering, the word bags after pre-processing were translated into a *TF-IDF* matrix (term frequency inverse document frequency matrix; [25]). A *TF-IDF* matrix consists of a term frequency part (*TF* part) and a normalizing factor which reduces the importance of very frequently occurring terms (*IDF* part). The inverse document frequency for the term i is defined as the logarithm of the total number of micro-texts T in the corpus divided by the number of texts n_i containing the term i:

$$IDF(i) = \log \frac{T}{n_i}, \tag{5}$$

where $i \in \{1, \ldots, I\}$ and I is the total number of terms. The element $v_{d,i}$ of the *TF-IDF* matrix V is then defined as

$$v_{d,i} = TF(i,d) * IDF(i), \tag{6}$$

where the term frequency $TF(i,d)$ denotes how often the term i occurs in text d.

Our clustering approach is an on-line version of latent topic clustering based on a principal component analysis (PCA) which is closely related to the well-established latent semantic indexing method [26]. The general aim of such methods is to uncover the underlying semantic structure of a collection of texts by finding latent topics, where – in the statistical versions – a topic is understood as a distribution over terms and each text is associated with different topics to different degrees. A simple classification can then be reached by assigning each text to its predominant topic. In the PCA approach, latent topics are basically represented by principal components in the space that contains the vectors $v_d = (v_{d,1}, \ldots v_{d,I})$, $d \in \{1, \ldots, T\}$.

The basic idea of the algorithm can then be summarized as follows:

1. Perform a singular value decomposition of the centered *TF-IDF* matrix V, which is equivalent to a PCA and involves an eigenvalue decomposition of the matrix $V^T V$.

2. Create a lower-dimensional representation of the data based on the k largest eigenvalues.

3. Assign each micro-text d to the principal component that has the biggest absolute coordinate value (scores) for the corresponding vector v_d. This yields the k basic topical clusters.

4. Each topic can be characterized by the corresponding eigenvector which, in turn, is characterized by choosing the most predominant base vectors (loadings), i.e. terms that determine the eigenvector.

This basic algorithm has the drawback that it does not allow for a straightforward integration of new data – neither of new dimensions (terms) nor of new items (micro-texts) and the decomposition must be recalculated in order to adapt to new data. V is a $T \times I$ matrix, where T and I are growing in time and typically $T \ll I$. Using standard algorithms, the time complexity for the decomposition is of order $O(I^3 + I^2T)$ for each recalculation [27].

To overcome the high computational costs and to enable online learning with a smooth adaption to new input we implemented a neural network-based variant of PCA, so-called Generalized Hebbian Learning; [28]. It is guaranteed to converge towards a standard PCA solution [29]. Here, we provide an outline of the generalized Hebbian algorithm (GHA), for further details see [28].

1. To start, we expect $T \geq 10$ for the learning procedure.

2. Initialize a neural network with I input neurons and k output neurons where I corresponds to the current number of terms and k is the number of classes chosen.

3. Initialize the synaptic weights $\omega_{j,i}$ to small random values with $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, I$. Assign a small value to the learning rate $\eta \leq 1$ and set the counter $n = 1$.

4. Choose an input vector x randomly from the data corresponding to a row v_d of the *TF-IDF* matrix.

5. Calculate the output according to

$$y_j(n) = \sum_{i=1}^{I} \omega_{ji}(n) x_i(n) \tag{7}$$

   Where $y_j(n)$ denotes the value of output neuron j and $x_i(n)$ the input to neuron i.

6. Calculate $\Delta\omega$ according to the following (Hebbian) learning rule

$$\Delta\omega_{ji}(n) = \eta[y_j(n)x_i(n) - y_j(n)\sum_{l=1}^{j} \omega_{li}(n)y_l(n)] \tag{8}$$

7. Adapt the synaptic weights ω_{ji} according to

$$\omega_{ji}(n+1) = \omega_{ji}(n) + \Delta\omega_{ji}(n) \tag{9}$$

8. Set $n = n + 1$ and repeat steps 4 to 7 until convergence.

After convergence, the synaptic weights of output neuron j represent the j-th principal component. Using GHA the system becomes adaptive to both increasing number of micro-texts and increasing number of features. As soon as a new text introduces a new term to the text corpora, a new input neuron is being added to the network and therefore the network adapts to the new feature space. The exact time complexity of the GHA algorithm depends on the accuracy needed. We found that with $k \ll T \ll I$, the leading order does not exceed $O(I^2)$ if the algorithm is terminated after at most I steps. Details regarding the choice of k are outlined in [28].

3.4 Top-down classification

The benchmark clustering of each project allowed creating an ideal classifier as follows: For each term of the 100 texts of a single project, we analyzed its relative frequency within each group of the group's benchmark. Those words that are specific only for one group form the classifier for each group. Due to the high variability among the groups and projects, we did not use a universal cutoff-value, but used individual adaptations for each project/group. In that sense, a classifier for a single project is defined by four word-bags, where each bag was characteristic for one group. The content of the word bag was adapted to the preprocessing and enrichment technique used, e.g., it contained also the English and French translations of a specific word when the classifier was applied to the texts of the corresponding step.

When performing the classification task, each text was compared to the classifier by calculating the relative overlap of the word bag of the text with each of the four word bags of the classifier, i.e. the number of words that are contained in both sets divided by the size of the smaller set. In this way, each text is represented by a 4-dimensional vector with coordinates between 0 and 1. Classification has then been achieved using the clustering algorithm of Mathematica – an adaptation of k-Means – for $k = 4$.

3.5 Combined bottom-up and top-down classification

We then have implemented the procedure discussed and outlined later in Fig. 3 that combines bottom-up clustering and top-down classification as follows. For all three projects, we used the preprocessing and enrichment procedure that lead to the most optimal results in the mean (see Section 3.1): word splitting & lemmatizing & translation. After preprocessing, the algorithm was performed as follows:

1. All texts have been clustered using Hebbian principal component clustering as outlined in Section 3.3 ($k = 5$).

2. For each identified cluster, we calculated the mean Euclidean distance between the texts that belong to this cluster on the basis of the rows v_d of the *TF-IDF* matrix.

3. We identified the cluster that achieved the smallest intra-group distances among cluster texts. The distribution of the intra-group distances had to be significantly different compared to the distances between all texts (Mann-Whitney test, $ps < 0.05$). This cluster has been identified as the best, i.e., the most distinguished cluster.

4. For all words of the cluster identified in step 3, we calculated their relative frequency within the cluster compared to their general relative frequency. All words with values exceeding 1 – i.e. they are more frequent in the cluster identified compared to their general frequency – were chosen as preliminary key words to characterize the identified cluster. This approach is slightly simpler than the description of cluster topics via eigenvectors (see 3.3). This simplified procedure can be justified by the next step that makes use of additional external intelligence.

5. The human supervisor decided which words should form the classifier $\{c_1, \dots c_l\}$ based on the selection of words that resulted from step 4.

6. For each text, we counted how often a word c_i of the classifier appeared. Let $\bar{n}(c_1, \dots c_l)$ be the mean number of how often all words of the classifier appeared per text and let $n_i(c_1, \dots c_l)$ be the number that denotes how often all words of the classifier appeared in text i. If $n_i(c_1, \dots c_l) > \bar{n}(c_1, \dots c_l)$, then the text i belongs to the identified group.

7. After removing the texts of the identified cluster, the procedure was repeated twice restarting from step 1.

This procedure identifies four clusters; three using the algorithm, the fourth cluster consisted of the remaining texts. We remark that this procedure is slightly adapted to the general problem outlined in Section 2, as our set of texts did not grow in time. In this way, the result was comparable to bottom-

up and top-down alone as well as to the human bottom-up experiment that we present in the next section.

3.6 Human bottom-up classification experiment

In the human bottom-up classification experiment, subjects rated the pairwise similarity of two ideas of a single project on a 7-point Likert scale. Each subject rated 30 pairs. We aimed for five ratings per pair, so that in total almost 25,000 ratings per project were required. Due to the large number of subjects needed for such an experiment, we used *Amazon Mechanical Turk* for recruiting participants (see `https://www.mturk.com/mturk/welcome`). *Amazon Mechanical Turk* is increasingly popular when performing psychological experiment and has been shown to provide reliable results [30]. We translated all texts to English for this task. All subjects received information on the background of the study and gave informed consent. This study was cleared in accordance with the ethical review processes of the University of Zurich and within the "Ethical Guidelines for Psychologists of the Swiss Society for Psychology".

We performed the following tests to exclude careless raters. First, each run included a test pair of ideas that were clearly similar – everybody who did not rate this pair accordingly was excluded. Second, we also excluded all participants that showed repetitive answering patterns independent of the idea pairs. After this check, the data of 875 subjects in the first project (exclusion rate 14.8%), of 901 subjects in the second project (exclusion rate 18.8%), and of 895 subjects in the third project (exclusion rate 13.4%) have been included for the analysis.

Based on the resulting rating data, we calculated the mean similarity for each pair and transformed the result in a normalized pairwise distance. Each text was then represented as a vector with coordinates between 0 and 1. Classification has then been achieved using the clustering algorithm of Mathematica for $k = 4$.

4. Results

4.1 Identification of optimal preprocessing and enrichment

We first report the result of bottom-up and top-down clustering for all 12 types of preprocessing and enrichment procedures. For each project and each type, we calculated the Jaccard coefficient both for bottom-up clustering as described in Section 3.3 and top-down classification according to Section 3.4 (6 data points in total) and we then calculated the mean improvement of the types 2 to 12 compared to type 1. We used Mathematica version 9 for data processing and general statistical analysis. Table IV shows the results.

We find that the effect of preprocessing and enrichment is quite variable and strongly dependent on

Table IV. Clustering improvement for text preprocessing types 2 to 12 compared to baseline (type 1) for both top-down and bottom-up-clustering. *Abbreviations: S: Stemming; T: Translation; H: word splitting, L: lemmatizing, Y: synonym enrichment. JC: Jaccard coefficient. **In the row "test statistics", it is indicated to which other preprocessing types a specified preprocessing type achieved significant improvements. $^a : p < .05$, $^b : p < .1$, n.s.: not significant.

Type	Abbreviation*	Mean JC improvement	Test Statistic (Mann-Whitney)**
2	S	.042	n.s.
3	T	.063	n.s.
4	H	.030	n.s.
5	H & T	.050	n.s.
6	H & L	.001	7^b
7	H & L & T	.143	6^b, 8^b, 9^a
8	H & L & Y	-.004	7^b
9	H & L & Y & T	-.046	7^a, 11^b
10	H & S	.052	n.s.
11	H & L & S	.081	9^b
12	H & L & Y & S	.028	n.s.

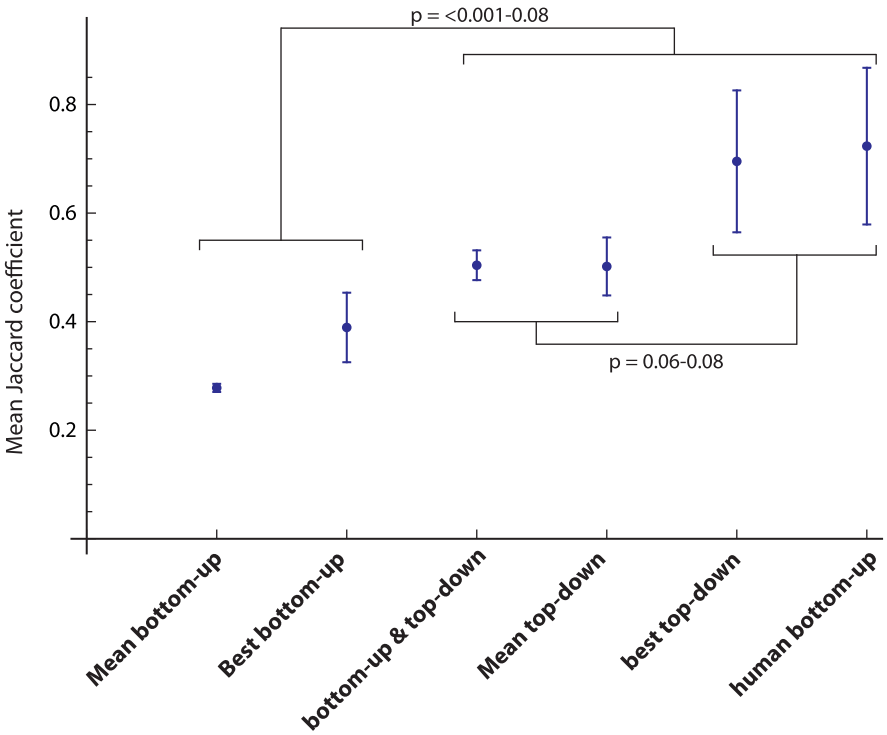

Fig. 2. Comparing the classification results bottom-up, top-down and the combination according to the protocol of Fig. 3 with human bottom-up classification.

the kind of texts, i.e. are related to the project. Because of this variability, only few preprocessing types achieve significant improvement. The result shows that translation, which has been introduced as an alternative to synonyms for text enrichment [23], is a quite powerful enrichment technique, whereas enriching texts by synonyms generally increases the similarity of all texts and thus worsen cluster discrimination: For example, when comparing type 7 with type 9 preprocessing, or type 11 with type 12 preprocessing, adding synonym enrichment significantly decreases the improvement – or makes it even worse than clustering of the original texts in type 9, where synonym enrichment and translation (after enrichment) have been combined. This indicates that translation after synonym enrichment even increases "blurring" of clusters. In the mean, the most successful type is the combination of word splitting, lemmatizing, and translation (type 7), by which up to 70% improvement was possible. This type also showed most often significant or tendencies for significant differences to the other types.

4.2 Comparison of the classification procedures

When comparing the results of the clustering approaches for all three projects, we find that the combination of bottom-up generated classifier with top-down classification is equally good as the mean result of all top-down classifications over all preprocessing and enrichment procedures when using the optimal top-down classifier (Fig. 2).

As we have found that the result of text preprocessing and enrichment is strongly dependent on the type of texts – i.e. it cannot be known a priori when the grouping of the texts is unknown – we can conclude that the combination achieves an optimal result with rather minimal intervention by a human supervisor. The result of bottom-up clustering measured by the Jaccard index is only about half as good and even the mean of the best bottom-up results over all three projects is clearly worse than our semi-supervised learning system.

An interesting result is that human bottom-up clustering is comparable to the mean of the best top-down classifications using an optimal classifier, although no subject had a holistic overview of all texts that had to be classified.

5. Discussion and conclusions

In this article we investigated which text classification approaches are best suited for medium-sized data-sets of micro-texts. We compared a purely bottom-up to a top-down approach which involve several types of text pre-processing and enrichment as well as a semi-supervised learning procedure (outlined in Fig. 3). We propose a new semi-supervised procedure that combines an unsupervised learning step to identify the best cluster(s) with a supervised control step to define a classifier.

The process for this new procedure includes three steps and is summarized as follows:

1. Preprocessing (stop word elimination, word splitting, lemmatizing) and text enrichment (translation) allows to increase the information content of the micro-texts for the following bottom-up classification procedure

2. The bottom-up classification involves a latent topic approach ("Hebbian clustering") that yields clusters and a corresponding characterization by means of key words. The step also yields an assessment of the clusters in terms of their "clusterness"; i.e., the quality of the clusters is evaluated.

3. The top-down procedure comprises two stages. First, a human supervisor intervenes by assessing the quality of the best cluster as suggested by the clustering algorithm and by confirming or

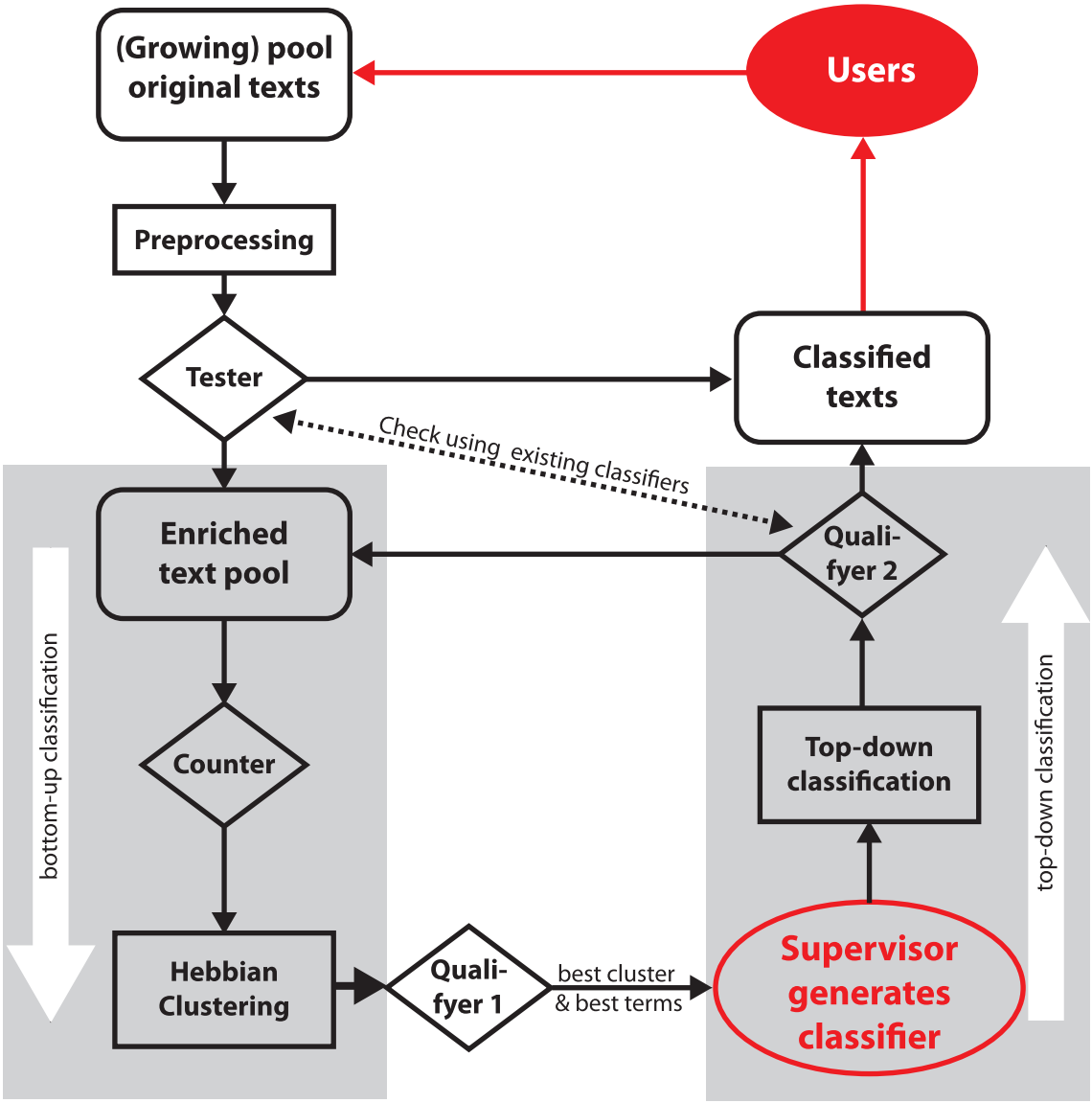

Fig. 3. Protocol of the enrichment, clustering and classification procedure of micro-texts.

changing the most significant key words that have been found for this cluster. Second, the identified key words are used to define an automated classifier for the corresponding best cluster. This classifier allows assigning new micro-texts directly if they match to the corresponding topic. If new data is entering the system or significant new clusters can be found, the entire procedure of clustering and classification is repeated.

We have shown for three different text content classes related to museums and art, to different types of food, and to marketing of a beverage that our protocol achieves an optimal classification result compared to the investigated alternatives given the fact that the kind and number of classes is unknown and that no pretesting for finding the optimal text enrichment can be done in real-time classification. We remind that the benchmark classification (section 3.1) is based on a consensus for a classification requiring the agreement of 8 out of 10 coders. Hence the benchmark classification itself is prone to ambiguity. For example, some texts of contest 2 discuss both the topic "soups" as well as the topic "heating" and both class assignments could be meaningful. So we can assume that about 5-10% of the texts could have been classified differently in the benchmark dataset. If you compare two classifications with 2 clusters (e.g. n = 50 points each), where 5-10% of the data items are missclassified, you end up with a Jaccard coefficient between about 0.7-0.84. We thus have to conclude that, even though the exact Jaccard coefficient depends on the actual clustering configuration, a classification result yielding a Jaccard coefficient between about 0.6 and 0.8 as in our case has to be considered a reasonably good result.

We also have found that human bottom-up classification – i.e. pairwise text similarity rating without a holistic view on the whole text set – reveals very good results, indicating that human input is needed in cases where the number of texts is insufficient for efficient machine learning. The reason for this may be that humans are able to grasp the context-dependent semantic relations among text also when only a low number of texts are available. We suspect that a regress on generalized semantic data bases for a completely automatized classification (e.g. *WordNet*, see `http://wordnet.princeton.edu/`) would not solve this problem, as large data bases are not context sensitive and the comparable low number of texts is unlikely to reveal this context sensitivity based on a purely automatic procedure. However, this supposition remains untested, as this is beyond the scope of our paper, as an elaborated semantic database like *WordNet* is not publicly available for the German language.

Our findings have implications for the management of internet platforms that contain medium-sized amounts of micro-text. Our proposed procedure allows clustering the content on these websites in a meaningful way requiring only little human intervention. Classification can make sense as it allows getting an overview of the content. On *Atizo.com* for instance, providing an overview over the generated ideas is beneficial both to the seekers and solvers. Solvers can more quickly observe which ideas have already been proposed which can reduce the number of redundant ideas. Seekers benefit by the lower processing time required to go through the ideas. Similar benefits might be achieved in any other site containing micro-texts in medium-sized amounts.

We remind some important shortcomings of our study. First, the optimal enrichment techniques depend on the type of content the micro-texts express. It is thus possible that for micro-texts of a different domain like, e.g., comments on social media platforms instead of idea drafts, another combination of preprocessing and enrichment procedures may reveal better results – in particular in case of other languages than German, where combined words are less frequent and word splitting may be less important. Therefore, the practical application of our protocol should be pre-tested for some known text clustering. However, we believe that translation remains an interesting enrichment technique as it is more context sensitive than semantic enrichment. Second, we did not systematically check other clustering techniques for the bottom-up part, although pre-tests [28] indicated that our approach achieves better results compared to k-means clustering. Finally, we did not perform a sophisticated comparison of this technique with other methods of information retrieval like supervised latent Dirichlet allocation [31] or other supervised topic models. In this study, we focused on outlining the procedure. Systematic comparison studies for a broader set of text categories and across several languages will be addressed in future work.

In summary, we have shown that for classification of intermediate sized sets of micro-texts, a "machine suggestion" and a human intervention that introduces a kind of nonlinear correction may be the optimal strategy. In cases when the amount of text is too small for revealing context and an appropriate text statistics, the human factor remains important for practical results.

Acknowledgments

We thank Reto Aebersold from *Atizo AG* (Bern, Switzerland) and Sara Irina Fabrikant (Geographic Information Visualization & Analysis, Department of Geography, University of Zurich, Switzerland) for their support in this research project. This project has been supported by the Swiss Commission for Technology and Innovation, grant 12747.1 PFES-ES.

References

[1] D. Helbing and S. Balietti, "From social data mining to forecasting socio-economic crises," *European Physics Journal - Special Topics*, vol. 195, pp. 3–68, 2011.

[2] S.K. Shriver, H.S. Nair, and R. Hofstetter, "Social ties and user-generated content: Evidence from an online social network," *Management Science*, vol. 59, no. 6, pp. 1425–1443, 2013.

[3] U.-D. Reips and P. Garaizar, "Mining twitter: A source for psychological wisdom of the crowds," *Behavioral Research*, vol. 43, pp. 635–642, 2011.

[4] R.E. Wilson, S.D. Gosling, and L.T. Graham, "A review of facebook research in the social sciences," *Perspectives on Psychological Science*, vol. 7, pp. 203–220, 2012.

[5] R. Tinati, S. Halford, L. Carr, and C. Pope, "Big Data: Methodological challenges and approaches for sociological analysis," *Sociology*, vol. 48, no. 4, pp. 663–681, 2014.

[6] C. Beghtol, "From the universe of knowledge to the universe of concepts: The structural revolution in classification for information retrieval," *Axiomathes*, vol. 18, pp. 131–144, 2008.

[7] K. Rosa and J. Ellen, "Text classification methodologies applied to micro-text in military chat," in *Proceedings of the International Conference on Machine Learning and Applications, ICMLA 09*, December 13-15 2009, Miami, Florida, USA, pp. 710–714, 2009.

[8] F. Peréz, D. Pinto, J. Cardiff, and P. Rosso, "On the difficulty of clustering microblog texts for online reputation management," in *Proceedings of the ACL-HLT 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA-2011*, June 24, Portland, Oregon, USA, 2011.

[9] A.H. Wang, "Don't follow me – spam detection in twitter," in *Proceedings of the International Conference on Security and Cryptography, SECRYPT 2010*, July 26-28, 2010, Athens, Greece, pp. 142–151, 2010.

[10] M. Salathé, C.C. Freifeld, S.R. Mekaru, A.F. Tomasulo, and J.S. Brownstein, "Influenza A (H7N9) and the importance of digital epidemiology," *New England Journal of Medicine*, vol. 369, pp. 401–404, 2013.

[11] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control," *PLoS Computational Biology*, vol. 7, no. 10, e1002199, 2011.

[12] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[13] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, May 23–26 2010, Washington, DC, 2010.

[14] C. Silva and B. Ribeiro, *Inductive inference for large scale text classification*, Springer, 2010, DOI: 10.1007/978-3-642-04533-2.

[15] M.D. Lee and E.Y. Corlett, "Sequential sampling models of human text classification," *Cognitive Science*, vol. 27, pp. 159–193, 2003.

[16] A. Cardoso-Cachopo and A.L. Oliveira, "An empirical comparison of text categorization methods," *Lecture Notes in Computer Science*, vol. 2857, pp. 183–196, 2003.

[17] M.K. Poetz and M. Schreier, "The value of crowdsourcing: can users really compete with professionals in generating new product ideas?," *Journal of Product Innovation Management*, vol. 29, no. 2, pp. 245–256, 2012.

[18] K. Girotra, C. Terwiesch, and K.T. Ulrich, "Idea generation and the quality of the best idea," *Management Science*, vol. 56, no. 4, pp. 591–605, 2010.

[19] R. Hofstetter, A. Herrmann, and J.Z. Zhang, "Incentives for crowdsourcing contests: Winner-takes-all or multiple prizes?," Working Paper, University of Lugano, 2015.

[20] L.J. Kornish and K.T. Ulrich, "Opportunity spaces in innovation: Empirical analysis of large samples of ideas," *Management Science*, vol. 57, no. 1, pp. 107–28, 2011.

[21] M. Popovi, D. Stein, and H. Ney, "Statistical machine translation of german compound words. Advances in natural language processing," in *Lecture Notes in Computer Science*, vol. 4139, pp. 616–624, Springer, Berlin, 2006.

[22] M. Hwang, C. Choi, and P. Kim, "Automatic enrichment of semantic relation network and its application to word sense disambiguation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 845–858, 2011.

[23] J. Tang, X. Wang, H. Gao, X. Hu, and H. Liu, "Enriching short text representation in microblog for clustering," *Frontiers of Computer Science in China*, vol. 6, no. 1, pp. 88–101, 2012.

[24] M.F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–147, 1980.

[25] D. Jurafsky and J.H. Martin, *Speech and language processing*. Prentice Hall, London, 2009.

[26] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[27] A. Sharma and K.K. Paliwal, "Fast principal component analysis using fixed-point algorithm," *Pattern Recognition Letters*, vol. 28, pp. 1151–1155, 2007.

[28] T. Niederberger, N. Stoop, M. Christen, and T. Ott, "Hebbian principal component clustering for information retrieval on a crowdsourcing platform," in *Proceedings of the 20th IEEE Workshop Nonlinear Dynamics of Electronic Systems, NDES-2012*, July 11-13 2012, Wolfenbttel, Germany, 2012.

[29] S. Haykin, *Neural networks. A comprehensive foundation*. Prentice Hall, London, 1999.

[30] M.D. Buhrmester, T. Kwang, and S.D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 3, no. 6, pp. 13–5, 2011.

[31] D.M. Blei and J. McAuliffe, "Supervised topic models," in *Proceedings of the 21. Annual Conference on Neural Information Processing Systems, NIPS 2007*, Vancouver, December 3-6, 2007.