A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests

Thomas P. Walter
Institute of Information Management,
University of St. Gallen
thomas.walter@unisg.ch

Andrea Back
Institute of Information Management,
University of St. Gallen
andrea.back@unisg.ch

Abstract

This survey deals with the problem of evaluating the submissions to crowdsourcing websites on which data is increasing rapidly in both volume and complexity. Usually expert committees are installed to rate submissions, select winners and adjust monetary rewards. Thus, with an increasing number of submissions, this process is getting more complex, time-consuming and hence expensive. In this paper we suggest following text mining methodology, foremost similarity measurements and clustering algorithms, to evaluate the quality of submissions to crowdsourcing contests semi-automatically. We evaluate our approach by comparing text mining based measurement of more than 40'000 submissions with the real-world decisions made by expert committees using Precision and Recall together with F_1-score.

1. Introduction

In 2006, Thomas Davenport argued in an article in Harvard Business Review that the latest strategic weapon for companies is analytical decision making, providing examples of companies that have used analytics to better understand their customers and optimize extended supply chains to maximize their return on investment while providing the best customer service [1]. A large component of this understanding comes from analyzing the vast amount of data that a company collects. The cost storing and processing data has decreased dramatically in the past, and, as a result, the amount of data stored in electronic form has grown at an explosive rate [2].

As mentioned in the call for paper of this minitrack, social media, encompassing a range of web sites such as blogs, microblogs, wikis, forums or social networks generate tremendous volumes of numerical and textual data that can be mined and analyzed for both research and commercial purposes.

This paper deals with the problem of analyzing data from crowdsourcing websites, a form of social media, on which data is increasing rapidly in both volume and complexity. Crowdsourcing websites, which often claim to tap a collective intelligence [3], [4] or the so called wisdom of crowds [5], have attracted worldwide attention of both, practice and the scientific community. In 2006, Jeff Howe defined crowdsourcing as the new pool of cheap labor: Everyday people using their spare cycles to create content, solve problems, even do corporate R&D, mostly by using crowdsourcing websites [6]. Whereas the basic idea behind the concept of crowdsourcing is rather clear, so far, neither crowdsourcing platforms nor research succeeded in submitting evidence which methods of measurement can or should be applied to analyze and evaluate submissions towards crowdsourcing websites. On the other hand, the return on investment of crowdsourcing is questioned by firms. For instance, firms, in quest of innovative product solutions via crowdsourcing websites, often obtain up to 1000 submissions, an amount that can be similar to 1000 pages of plain text data. Usually expert committees are installed to rate submissions, select winners and adjust monetary rewards. However they often are unable to cope with this sheer quantity and complexity of data. As a consequence, firms are running the risk of missing the benefits of crowdsourcing.

In this paper we suggest to follow a text mining approach to address the given problem. Text mining is the semi-automated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources [2]. Text mining works by transposing words and phrases in unstructured data, such as submissions to crowdsourcing websites, into numerical values which can then be linked with structured data in a database and analyzed with data mining techniques [7], [8]. Our goal is to provide decision support to the expert committees' process of analyzing and evaluating submissions to crowdsourcing websites. As it is a longstanding dream of the community to have algorithms that are capable of automatically reading and obtaining knowledge from text our initial research question is stated as following:

Accepted manuscript for 46th Hawaii International Conference on System Sciences (HICSS 2013).

RQ1: How can text mining methodology be applied to support the submission evaluation process on crowdsourcing websites by suggesting most innovative solutions?

For this purpose we provide theoretical background on crowdsourcing websites in general and current methods of crowdsourcing evaluation in particular during chapter 2. Furthermore we conduct a brief literature review on papers which deal with text mining approaches in crowdsourcing evaluation. Chapter 3 focuses on the development of the text mining approach itself. This includes the description of the applied text mining methods as well as the dataset we use to test our approach. We exploit platform data from a real crowdsourcing website. This data includes over 100 finished crowdsourcing contests together with all raw text data given by over 40'000 submissions. Furthermore we make use of real-world expert committees' decisions about those submissions, foremost which are most valid to seeking companies and hence, are rewarded. This enables us to state our second research question as following:

RQ2: To what extend can a text mining based evaluation of submissions to crowdsourcing websites reproduce the results of expert committees in regards to selecting most innovative submissions?

We present our answer to RQ1 during Chapter 3 and results to RQ2 in Chapter 4. Analysis of the data was performed using accuracy measures from the field of Information Retrieval, Precision, Recall and F_1-score. This enables us to compare results from the manual expert committees' decisions with the semi-automated, text mining based evaluation approaches. Chapter 5 aims on drawing conclusions from this survey, including managerial and theoretical impact as well as current limitations and an outlook to further studies.

2. Background

The increasing popularity of open innovation approaches [9] in practice has led to the rise of various literature streams within the area of crowdsourcing. Following, we will focus on three central aspects: how crowdsourcing success currently is defined, how submissions from the crowd are evaluated on crowdsourcing websites and to what extend the evaluation is already supported by text mining approaches.

2.1. Success Patterns of Crowdsourcing

Defining success patterns of crowdsourcing opens a two-sided discussion. On the one hand, various studies find positive effects of monetary rewards on quantity aspects of crowdsourcing, foremost the amount of attracted solvers or the amount of submissions [10–15]. In general firms can benefit from larger crowds because they obtain a more diverse set of solutions [3], [5], [16], [17], which mitigates and sometimes outweighs the effect of the crowds' underinvestment [18], [19]. Accordingly research states that it requires a large amount and variety of submissions to achieve a high quality best idea [19–22].

On the other hand, economists state that with a large-scaled crowd, each member will have relatively small chance of winning, so the winner's investment and hence, the quality of the winning submission will tend to be low [23], [24]. Furthermore large amounts of submissions slow down the evaluation process due to the necessity of filtering signals from noise. Finally, to the best of our knowledge, there is only little empirical evidence on what drives the quality of crowdsourcing. [25] find that highly connected crowds tend to produce lower quality, [21] find quality to be dependent on adequate crowd coordination techniques, [26] finds individual quality to be positively related to current effort, but negatively related to past success within crowdsourcing, [27] find that crowd performance rises after they recognize being above average, [28] find that, in comparison to experts, on average crowd submissions score higher in novelty and customer benefit, but lower in feasibility and [29] find that whether a task was framed as meaningful does not induce greater or higher quality output.

2.2. Measuring the Quality of Crowdsourcing

Next to these findings on general success patterns, research is unclear about how to measure and define the quality of crowdsourcing. As in [15] the size of the attracted crowd is taken as indirect measurement of quality, [14] use the scale of every submission on a five-star rating, [29] take the information whether a task was completed as a measurement and [26] uses the information whether a submission was eventually implemented as primary dependent measure of quality. Also qualitative approaches can be found. [25] use data from external experts to measure quality and [28] take the evaluation from independent executives to compare crowd and expert submissions.

In contrast, [30] find that all simple rating mechanisms, such as thumbs up/ down or 5-star ratings are not sufficient to measure the quality of submissions and suggest a multi-attribute scaling including ratings from both, independent expert committees and crowds. In the context of crowd-generated product ideas, [31] aggregate literature and define that idea quality consists of four distinct dimensions: novelty, feasibility, strategic relevance and elaboration. Novelty

typically is defined as something being unique, rare or not been expressed before [32]. Another attribute of novelty is the relatedness among submissions [33–35]. This refers to a revolutionary submissions character of being radical and not related to others. Closely related to novelty is originality. Originality of submissions can be defined by their ability to surprise, imaginariness or degree of unexpectedness [36]. Hence, following this Schumpeterian definition of innovation, many researchers see novelty and originality as the most important facet of creativity [30], [34], [36] and hence, as most suitable measurement of crowdsourcing quality.

2.3. Using Text Mining to Measure the Quality of Crowdsourcing

Literature applying text mining methodology is manifold and spread over diverse research fields. For instance, text mining has become an appreciated research methodology different research areas, from patent analysis [37] towards biology [38]. However, although crowdsourcing websites are generating tremendous volumes of numerical and textual data, a brief, but specific literature review of applied text mining methodology on crowdsourcing or the related area of collective intelligence does not provide a plurality of papers. Therefore, we scan an IS-specific database (The Association of Information Systems electronic Library, AISeL) using the search terms "text mining" and major topics crowdsourcing and collective intelligence combined by a logical AND. Findings are diffuse and the coverage of crowdsourcing websites can be described as rather vague. [39] apply four commonly used text classification algorithms and propose a text classification framework for finding helpful user- generated contents in online knowledge-sharing communities. [40] present and evaluate different manual, semi-automatic, and automatic text analysis methods for summarizing transcripts transforming tacit knowledge into explicit form and to substantially reduce the time required to perform this transformation. [41] run text mining methodology on user opinions, expressed via twitter to analyze the appearance of a collective intelligence. [42] develop a taxonomy for combining text and data mining. [43] use text mining to analyze different genres of spam and [44] apply text mining to depict crime networks. Table 1 summarizes the provided background literature.

The final row within Table 1 also represents the research gap we address with our survey. Text mining methodology is used to analyze various aspects of online communities, but to the best of our knowledge not yet to evaluate submissions to crowdsourcing websites.

Table 1. Summary of background literature

Research Stream	Representative Literature
Defining success patterns of crowdsourcing	[3], [5], [9–16], [18]
Measuring quality aspects of crowdsourcing	[14], [15], [21], [25], [26], [28]
Analyzing and defining metrics to measure the quality of submissions to crowdsourcing websites	[30–36]
Applying text mining on open online communities	[39–44]

3. Methodology

This study exploits text mining to analyze a real-world data sample from an international crowdsourcing website. For the semi-automated step of evaluating the quality of submissions by text mining, clustering is implemented and compared to real-world results, that is to say expert committee decisions. The background literature sets the focus on two central points which describe a current research gap.

- A submission that can be described as representative or average will seldom be able to convince a problem seeking firm as it typically is often not the richest in information. Under the given circumstances of crowdsourcing websites, it is more useful to select submissions that offer an interesting, unusual or particularly revealing set of circumstances, submissions which are outstanding.
- Yet, text mining is not applied to fulfill a semi-automated selection of submissions to crowdsourcing websites.

During this chapter we will describe our approach to fill this gap. For text mining procedures foremost *Provalis Researchs QDA Miner*, including the extension package *Wordstat* is used and for the statistical analysis *R* is used.

3.1. Dataset and descriptive Statistics

To apply a text mining based approach of analyzing crowd submissions we make use of crowdsourcing website data. The website was launched in 2008, currently has over 7.000 active members, called solvers. Since its launch, 112 crowdsourcing contests have been closed. In average a contest is open for two month. Firms (called seekers) use the website to state a problem or task and crowds participate by logging in and submitting ideas or concepts to contests. External

incentives of participation are monetary rewards and a community ranking of solvers, which is also based on earned total rewards. The quartile of rewards is US$ 92 to 350 per selected (winning, rewarded) submission, but the reward structure is defined by expert committees after a contest has been closed. Hence, the quartile of contest reward-budgets is set by US$ 2'532 to 4'211 which is used to price selected winning submissions per contest. On average 379 ideas or concepts are submitted per contest. As the average length of a submission is 25 words of plain text (which equals 7 sentences or 115 characters), this makes our raw data 42'448 submissions (or 3.65 Mio words). Table 2 summarizes the metrics of the website data.

Table 2 Metrics of submissions to the crowdsourcing website

Unit	Total	Avg.	Std.dev
Crowd Submissions	42'448	379.0	87.41
- by sentences	154'110	2'653.25	559.12
- by words	1'078'771	9'631.89	2'108.4

To make this numbers more feasible, one could say that per average crowdsourcing contest text, twice the length of this conference paper, including 379 more or less outstanding ideas or concepts is submitted, and has to be evaluated and rewarded by expert committees. Contests concern different areas such as product ideas, marketing concepts or technical solutions and are demanded by firms operating in various industries. The following text may serve as an example of a representative crowdsourcing contest. It is an excerpt from a task, provided by a global player in the sports apparel industry, offering a total of US$ 6'000 for the most innovative submissions:

"How can the clothing of the future better regulate the sportsman's body temperature? Athletes at the Olympic Games efforts were challenged by the high temperatures. Athletes produce heat but only around 20% of it is utilized as energy, around 80% literally becomes "hot air", if the body cannot get rid of this heat, and this can lead to cramps and even heat strokes. So that our body does not overheat, it uses four preventive measures: sweating (evaporation), fanning oneself (convection, ventilation), channeling extra heat (conduction), radiates heat. Other examples exist, albeit lavish ones: ice-vests to cool down before competition, integrated ventilation systems in clothing (Air Force pilots). How can we develop a simple piece of clothing, which utilizes the four mentioned mechanisms to prevent athletes from overheating? The Evaluation Criteria are a) A Product that has not yet been developed (originality) and b) Quantifiable temperature reduction (effectiveness)."

3.2. Pre-Processing Crowdsourcing website data

The data was processed following a standard text mining procedure, e.g. as in [7]. The first major step of text mining is pre-processing. After extracting all raw text data (the so called corpus) from the website using simple MySQL statements, the raw data is imported to the QDA Miner software. Figure 1 illustrates this by using the given example of the sports apparel contest.

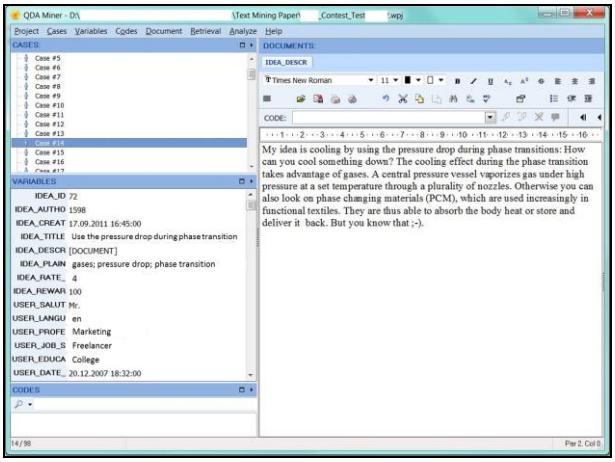

Figure 1 Typical answer to a crowdsourcing contest, shown as raw text data within the QDA Miner software.

This particular contest has one of the lowest amounts of submissions (98, depicted as cases in Figure 1). Nevertheless, the crowd submitted manifold types of answers to this contest, some from a technical focus, some from a rather simple minded focus. To be able to analyzed this data with traditional data mining techniques, text mining works by transposing words and phrases into numerical values which can then be linked with structured data in a database [7], [8].

Hence as a next step we run pre-processing steps, to be exact stemming, stop-word cleaning and tokenization. Stemming is the process for reducing inflected words to their stem, base or root form. For instance, stemming algorithms like [46] are used to delete suffixes. As a result a stemming algorithm would reduce words like computer, computing, and compute to their stem, which is "comput". Stop-word cleaning usually is partly a manual process. An algorithm searches text by a predefined list of so called stop-words and deletes them from the text. Most common, short function words, such as "the, is, at, but, which and on" are set onto stop-word lists [7]. Tokenization is the process of breaking a stream of text up into words, phrases or symbols and Part-of-speech tagging is the process of marking up a word in a text as

corresponding to a particular part of speech (to syntax) based on both its definition, as well as its context.

The term document matrix (TDM) is the final result of pre-processing. A TDM describes the frequency of terms which occur in a collection of text. In a TDM, rows correspond to documents (D) in the collection and columns correspond to terms (T). In our case documents are represented by submissions (called cases within Figure 1) to a specific crowdsourcing contest and terms are represented by words used within those submissions. Weighting of terms can be calculated binary (e.g. a certain expression is included in a collection), normalized (term frequency, tf) or by inverse term frequencies (tf-idf), which means overweighting less used terms within an collection purposely.

Following the background literature we use both, normalized (tf) and inverse term frequencies (tf-idf), to be able to compare results afterwards. Hence, we calculate two TDMs for each of the 112 crowdsourcing contests. In the enlightened case, the sports apparel contest, the TDM contains 194 x 89 data fields, stating 194 different words used at least in one of 98 submitted cases. However the largest TDM within our dataset is given by a 242 x 957 matrix, stating 242 different words within 957 submissions to one single contest.

3.3. Clustering Submissions to Crowdsourcing Contests

Text clustering is the application of certain algorithms to automatically detect patters within a TDM. Clustering is used to explore the similarity between documents. Often so called non-hierarchical (or centroid-based) clustering is applied, foremost the k-means algorithm [47–49]. In this survey k-means aims to partition text-documents (submissions) into k clusters in which each observation belongs to the cluster with the nearest mean. At the bottom k-means is based on principal component analysis or minimalizing least squares [50]. However, determining the k number of clusters in a data set is a frequent problem in data clustering, and is a distinct issue from the process of actually solving the clustering problem. In text mining, a frequently used method to determine the number of clusters can be estimated by the following formula (D x T) / t where t is defined as the amount of non-zero entries in the entire TDM [51].

Two broad types of clustering can be applied: first- and second-order clustering. First order clustering will group together words appearing in the same document and second order clustering will consider that two words are close to each other, not necessarily because they co-occur in the same document, but because they both occur in similar environments. One of the benefits of this clustering method is its ability to group words, and submissions therewith that are synonyms or alternate forms of the same word. For example, while TUMOR and TUMOUR will seldom or never occur together in the same document, second order clustering may find them to be related because they both co-occur with words like BRAIN or CANCER [52]. As a consequence we apply second order clustering.

Ultimately, clustering legitimizes a statement about the distance between all submissions within a contest. When the clustering is set to be performed on documents (submissions), a distance matrix used for clustering and multidimensional scaling consists of cosine coefficients computed on the relative term frequencies (in tf or tf-idf) of the various words within documents. The more similar two submissions will be in terms of the distribution of words, the higher will be this coefficient [7], [53]. Figure 2 illustrates the similarities within the given example of the sport apparel contest.

Cooccurrences | Similarity | Agglomeration

	Case #1	Case #2	Case #3	Case #4	Case #5	Case #6	Case #7	Case #8	Case #9	Case #10	Case #11	Case #12	Case #13	Case #14
Case #1	1.000													
Case #2	0.147	1.000												
Case #3	0.122	0.196	1.000											
Case #4	0.108	0.107	0.168	1.000										
Case #5	0.431	0.210	0.073	0.097	1.000									
Case #6	0.154	0.238	0.242	0.187	0.063	1.000								
Case #7	0.146	0.145	0.107	0.240	0.166	0.358	1.000							
Case #8	0.358	0.226	0.119	0.178	0.157	0.201	0.125	1.000						
Case #9	0.204	0.469	0.261	0.352	0.177	0.301	0.316	0.225	1.000					
Case #10	0.084	0.464	0.174	0.119	0.067	0.177	0.042	0.167	0.299	1.000				
Case #11	0.241	0.177	0.172	0.238	0.101	0.196	0.139	0.153	0.223	0.224	1.000			
Case #12	0.221	0.287	0.138	0.291	0.173	0.345	0.335	0.258	0.305	0.235	0.310	1.000		
Case #13	0.115	0.293	0.069	0.182	0.078	0.115	0.165	0.124	0.182	0.174	0.102	0.128	1.000	
Case #14	0.120	0.229	0.153	0.182	0.171	0.112	0.134	0.207	0.246	0.253	0.179	0.208	0.152	1.000
Case #15	0.110	0.255	0.269	0.233	0.140	0.255	0.162	0.156	0.245	0.130	0.223	0.223	0.150	0.225
Case #16	0.192	0.280	0.291	0.305	0.197	0.366	0.364	0.281	0.421	0.128	0.314	0.333	0.181	0.209
Case #17	0.174	0.189	0.115	0.174	0.127	0.247	0.212	0.301	0.252	0.191	0.219	0.335	0.118	0.174
Case #18	0.327	0.277	0.163	0.178	0.305	0.436	0.252	0.179	0.304	0.094	0.261	0.316	0.143	0.142
Case #19	0.189	0.315	0.200	0.365	0.213	0.270	0.156	0.308	0.263	0.133	0.268	0.311	0.193	0.227

Figure 2 Excerpt of similarity index matrix showing cosine coefficients between submissions to a crowdsourcing contest.

Each cosine coefficient is calculated by comparing term frequencies. Hence, the similarity of two submissions to a crowdsourcing contest will range from 0 to 1, since the term frequencies (tf or tf-idf weights) cannot be negative. The resulting similarities ranges from 1, meaning submissions are exactly the same (use exactly the same words) to 0, usually indicating a total independence, and in-between values indicating intermediate similarity or dissimilarity of submissions. In our example, depicted in Figure 2, this means submissions (called cases) #2 and #9 (in blue) to be much more similar than #5 and #6 (in red) for instance.

Hence, as a final step we have to select, which submissions should be selected within the text mining approach? Following the background literature (c.f. Table 1), an average, or typical submission is often not the richest in terms of novelty or originality. In other words, a selection that is based on representativeness will seldom be able to produce highly valuable insights for seeking firms. In clarifying lines of history and causation it is more useful to select submissions that offer an interesting, unusual or particularly revealing set of words. Following the literature on clustering, these kinds of submissions will stand out by their very unique set of used words. Hence, their cosine coefficients will be low towards most other submissions. Figure 3 illustrates an excerpt from the clustering in form of a dendrogram.

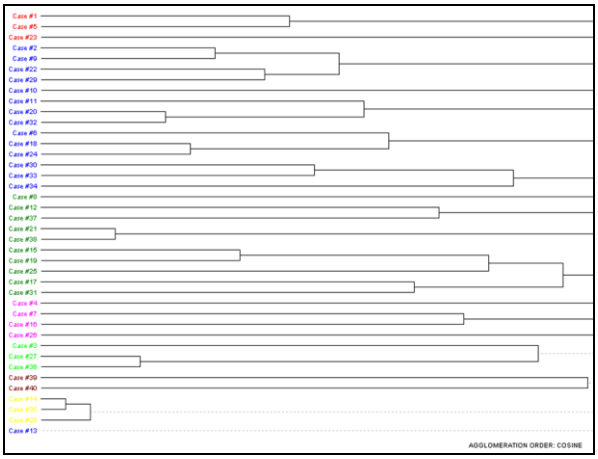

Figure 3 Dendrogram of clustered submissions within a crowdsourcing contest.

Clusters are visible by color, their aggregation is defined by their cosine coefficients, and the amount of clusters is calculated following the formula from [51]. Hence, there are different amount of clusters, including different amounts of submissions for each contest. As mentioned, following theory most innovative submissions should stand out, which means, at best they are not even part of a cluster at all (stating a so called single-item cluster). Therefore, we take two sets of submissions into the analysis, that is all submissions which appear as single document cluster, and second, all submissions which are part of clusters with up to three documents (submissions).

3.4. A Text Mining Based Evaluation of Submissions to Crowdsourcing Contests

We made use of theory to define what separates high quality submission from average submissions and we applied text mining methodology for semi-automated detection of submissions which supposed to be of high quality. Table 3 summarizes the process in analogy to the standard text mining process depicted in [7] and gives an answer to our first research question: *How can text mining methodology be applied to support the submission evaluation process on crowdsourcing websites by suggesting most innovative solutions?*

Table 3 Process of a text mining based evaluation of crowdsourcing contests.

	Process-Step	Method	Results
1	**Data extraction**	SQL-statements on website database	Raw-text data: 112 contest with 42'448 submissions
2	**Pre-Processing**	Stemming, stop-word cleaning and tokenization	Cleared data set
3	**Term Document Matrix (TDM)**	Two weighting algorithms: tf and tf-idf	Term frequencies in all contest, calculated by submissions
4	**Text Mining**	Calculating similarity (cosine coefficients) and clustering (k-means)	Clustered submissions per contest. (similar submissions as cluster)
5	**Submission selection**	Defining single-case- and cluster containing two or three submissions as outstanding	Semi-automated selection of best submissions

After raw text data is extracted from the crowdsourcing website, pre-processing is used to clear the data from meaningless terms and preparing it for text mining procedures. For each crowdsourcing contest, the TDM is calculated as words (terms, T) by submissions (documents, D). Overweighting less used terms by using the tf-idf format may already highlight outstanding ideas. Calculating the similarity of submissions within one contest by using their cosine coefficients opens the possibility to aggregate submissions into cluster. Following literature, we define cluster which include only one, or a maximum of three submissions, to contain ideas of outstanding quality and hence, the submissions which are selected to be rewarded.

4. Results

To evaluate the text mining approach, described during chapter 3, we measure its overall accuracy. Therefore we compare the two given kinds of selection processes, the real-world decisions by expert committees against the semi-automated submission selection process applying the described text mining

approach. The intention is to answer our second research question: *To what extend can a text mining based analysis of submissions to crowdsourcing websites reproduce the results of expert committees in regards to selecting most innovative submissions?*

Hence, the simple overall model to test is whether the text mining based approach is capable of reproducing the expert results. This makes text mining based selections our independent variable and the real world expert committee decision our dependent variable. Still, as we use different methods of measurement during the text mining approach, four different models have to be evaluated.

- Model A uses a TDM of type tf and clusters with only one submission to define which submissions are selected.
- Model B uses TDM of type tf-idf and clusters with only one submission.
- Model C uses TDM of type tf and clusters with up to three submissions.
- Model D uses TDM of type tf-idf and clusters with up to three submissions.

Following, all 42'448 submissions are used to evaluate those models. Evaluation follows standardized measurements from the field of Information Retrieval [54]. A descriptive analysis of all four models in terms of the selection task is shown in Table 4. The four quadrants of the so called confusion matrix [55] exhibit the absolute values of classifications made by the text mining approach, i.e. true positive results at top left, false positive at top right, false negative at bottom left ad true negative at bottom right. For instance, using model A a total of 635 submissions are selected. 522 of them are true positive, i.e. selected by both, the text mining approach and expert committees. Therefore, these are so called "hits".

Table 4 Selection of submissions made by the text mining approach compared to expert decisions.

		model	Selected by expert committees	
			Yes	**No**
Selected by applying text mining (models)	**Yes**	A	522	113
		B	367	69
		C	1'222	351
		D	949	237
	No	A	1'798	40'085
		B	1'883	40'129
		C	1'028	39'887
		D	1'305	39'957

However, at the same time model A produces 1'798 false negatives, which are called "misses". These submissions are only selected by the expert committees and not found by the text mining approach. On the other hand, 113 false positive submissions are only selected by the text mining approach, but neglected by experts. Finally, the vast amounts of submissions are true negatives, stating not being selected in any of the two ways. In a next step the absolute values from confusion matrix are used to calculate common metrics which measure the accuracy of the text mining approach, that is to say Precision, Recall and F_1-score. Those metrics are calculated as following [54]:

- Precision = true positives / (true positives + false positives)
- Recall = true positives / (true positives + false negatives)
- F_1-score = 2 * Precision * Recall / (Precision + Recall)

Table 5 summarizes those metrics for all four models. The results show that all models score rather high in precision and rather low in recall. This means all models tend to have a low amount of false positives, but unfortunately also a rather high amount of false negatives. In other words, selected submissions are mostly included in the experts picks, but experts mostly pick further submissions on top. That also explains higher Recall and F_1-score values for models C and D. As those models include cluster including up to three submissions, they simply select more submissions, which comes closer to the behavior of the experts.

Table 5 Precision, Recall and F_1-Scores for all models.

model	**Precision**	**Recall**	**F_1-score**
A	82.2 %	23.2 %	0.362
B	84.1 %	16.3 %	0.273
C	77.9 %	54.3 %	0.639
D	80.0 %	42.2 %	0.552

By not overloading rare terms, that is using the tf instead of the tf-idf type of a TDM, and not limiting selections to single submission clusters only, Model C is most valid in reproducing the decisions from various expert committees (F_1-score of 0.639), mostly because is scores highest in Recall. The inverse relation between Precision and Recall can be described as rather typical. For instance, one can often increase Recall by simply retrieving more documents [54]. In our case this would be possible by expanding the applied cluster sizes within the models. Ultimately, those results are a direct consequence to our initial definition of quality. The results show that in terms of used words, expert committees are also rewarding standard or average submissions. This does not mean that outstanding ideas get lost. In fact, high Precision in

all models shows that using unique sets of words correlates with the chance of a submission of being selected. But the results also show that this aspect does not do the entire trick of evaluating submissions.

6. Discussion and Conclusion

As stated, it is a longstanding dream of the community to have algorithms that are capable of automatically reading and obtaining knowledge from text, that are capable to understand human language. Despite great achievements in the field of text mining and natural language processing [7], [49], [53], [56], we will not have such possibilities in the near future. As stated in [58], many researchers think it will require a full simulation of how the mind works before we can write programs that read and understand the way people do.

So what can we learn from our study? We used long existing text mining algorithms and applied them on the modern research field of crowdsourcing contests. Our intention was to detect outstanding, innovative ideas, submitted by crowds, due to their likelihood of using unique sets of words and hence, separating them from a mass of so called noise. The empirical results are based on over 40'000 submissions. They show that text mining can serve as an approach to detect outstanding ideas. However, our approach has shortcomings and hence, should rather be seen as an initial step.

Overall all four models can be described as rather conservative selectors [54], meaning very few documents get selected in general and this is causing rather low Recall scores. This is due to the fact that following literature on crowdsourcing quality, we intended to focus on uniqueness. In contrast, expert committees seem to rather give plenty of lower rewards than following a "winner takes it all" strategy, which is a slightly different definition of quality. Hence, when it comes to selection of winning submissions, a text mining based approach must also be aligned with the reward structure of a crowdsourcing platform. Additionally, we treated all contests the same in regards to the semi-automated selection process. However, the 112 different contests addressed different topics, had different expert committees who applied individual reward structures and had differentiating opinion about relevant measuring criteria, especially the weighting of a submissions' uniqueness. Though it is our belief that taking those criteria into account would improve the results, it also goes beyond the scope of this paper, i.e. taking an initial step. A final shortcoming to mention is that we clustered submissions by their common use of unique words and therewith neglected the possibility of using more sophisticated clustering, e.g. by n-gramms or phrases. Future research should also elaborate on the question, which clustering method works best for comparing submissions to crowdsourcing platforms.

To sum it up, we do not suggest that a complete evaluation process should be based on text mining. Text mining could be used as decision support of expert committees as it provides fast and direct entrance to unique ideas. Concerning the rising problem of an increasing number of ideas, concepts or solutions being submitted by the crowd, text mining could facilitate the current situation of which expert committees commonly are unable to cope with. A final example should give further evidence to this result. Within our used example of the sports apparel contest the following idea (in raw text) received the highest reward from the expert committee:

"You could embed super-absorbent polymers (SAPs) into your clothing. The garment is now able to produce several types of cold: Total cool down (SAPs with cold water), cool down (SAPs with water at body temperature, cooling by evaporation), warming (SAPs with warm water). I don't know SAPs by hard, eventually evaporation of water from SAPs is too slow in your case and […]."

As in this case the expert committee had to read through 98 submissions to detect this submission, applying the text mining approach it becomes visible on first sight, mainly because of the uniqueness of the term "SAPs" within this contest. Hence, within the text mining approach, the same submission had the lowest average cosine coefficient, which means it had fewest in common with the other ideas, and therefore was selected by all four models. Still it can and should not be concluded from a study of a single data sample of the applied text mining approaches, which combination or combinations of these should be implemented in any particular situation. However, the identification in this study may guide future efforts to determine ideal combination of text mining algorithms during the evaluation of crowdsourcing contests.

7. References

[1] T. H. Davenport and J. G. Harris, Competing on Analytics: The New Science of Winning. Harvard Business School Publishing, 2007, p. 218.

[2] E. Turban, R. Sharda, and D. Delen, Decision Support and Business Intelligence Systems. Prentice Hall, 2011, p. 696.

[3] T. W. Malone, R. Laubacher, and C. Dellarocas, Harnessing Crowds : Mapping the Genome of Collective Intelligence. MIT Sloan School of Management Working Paper No. 4732-0, pp. 1-20 pp, 2009.

[4] E. Bonabeau, Decisions 2.0 : The Power of Collective Intelligence, MIT Sloan Management Review, vol. 50, no. 02, pp. 45-52, 2009.

[5] J. Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Little Brown, 2004.

[6] J. Howe, The rise of Crowdsourcing, Wired Magazine, vol. 14, no 6, pp. 1-4, 2006.

[7] R. Feldmann and J. Sanger, The text mining handbook : advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.

[8] Y. Kano et al., U-Compare: share and compare text mining tools with UIMA, Bioinformatics (Oxford, England), vol. 25, no. 15, pp. 1997-8, Aug. 2009.

[9] H. W. Chesbrough, Open innovation: The new imperative for creating and profiting from technology. Harvard Business School Press, 2003.

[10] M. Ahonen and M. Antikainen, Supporting collective creativity within open innovation, Proceedings from the 7th European Academy of Management conference (EURAM), pp. 1-18, 2007.

[11] M. J. Antikainen and H. K. Vaataja, Rewarding in open innovation communities – How to motivate members?, International Journal of Entrepreneurship and Innovation Management, vol. 11, no. 4, pp. 440 - 456, 2010.

[12] K. R. Lakhani and J. A. Panetta, The Principles of Distributed Innovation, Innovations: Technology, Governance, Globalization, vol. 2, no. 3, pp. 97-112, 2007.

[13] O. Stewart, D. Lubensky, and J. M. Huerta, Crowdsourcing Participation Inequality : A SCOUT Model for the Enterprise Domain, Proceedings from KDD-HCOMP'10, pp. 30-33, 2010.

[14] T. P. Walter and A. Back, Towards Measuring Crowdsourcing Success: An Empirical Study on Effects of External Factors in Online Idea Contests, Proceedings from the 6th Mediterranean Conference on Information Systems (MCIS), pp. 1-12, 2011.

[15] Y. Yang, P.-Y. Chen, and P. Pavlou, Open Innovation : An Empirical Study of Online Contests, Proceedings from the 30th International Conference on Information Systems (ICIS), pp. 1-16, 2009.

[16] J. M. Leimeister, Collective Intelligence, Business & Information Systems Engineering (BISE), vol. 02, no. 04, pp. 245-248, 2010.

[17] C. Wagner and A. Back, Group Wisdom Support Systems: Aggregating the Insights of Many through Information," Information Systems, vol. IX, no. 2, pp. 343-350, 2008.

[18] U. Gneezy and A. Rustichini, Pay Enough or Don't Pay at All*, Quarterly Journal of Economics, vol. 115, no. 3, pp. 791-810, 2000.

[19] C. Terwiesch and Y. Xu, Innovation Contests, Open Innovation, and Multiagent Problem Solving, Management Science, vol. 54, no. 9, pp. 1529-1543, 2008.

[20] 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.

[21] A. Kittur and R. E. Kraut, Harnessing the wisdom of crowds in wikipedia, Proceedings of the ACM 2008 conference on Computer supported cooperative work CSCW 08, p. 37-47, 2008.

[22] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, Cheap and Fast---But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254-263, 2008.

[23] N. Archak and A. Sundararajan, Optimal Design of Crowdsourcing Contests, , Proceedings from the 30th International Conference on Information Systems (ICIS), pp. 1-16, 2009.

[24] Y.-K. Che and I. Gale, Optimal Design of Research Contests, The American Economic Review, vol. 93, no. 3, pp. 646-671, 2003.

[25] J. Björk and M. Magnusson, Where Do Good Innovation Ideas Come From? Exploring the Influence of Network Connectivity on Innovation Idea Quality, Journal of Product Innovation Management, vol. 26, no. 6, pp. 662-670, 2009.

[26] B. L. Bayus, "Crowdsourcing and Individual Creativity over Time: The Detrimental Effects of Past Success," Working Paper of the Wharton Interactive Media Initiative, pp.1-34, 2010.

[27] B. A. Huberman, D. M. Romero, and F. Wu, Crowdsourcing , Attention and Productivity, Journal of Information Science, vol. 35, no. 6, pp. 758-778, 2009.

[28] 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., no., pp. 1-37, 2009.

[29] D. Chandler and A. Kapelner, Breaking monotony with meaning: Motivation in crowdsourcing markets, University of Chicago mimeo. 2010.

[30] C. Riedl, I. Blohm, J. M. Leimeister, and H. Krcmar, "Rating Scales For Collective Intelligence in Innovation Communities: Why Quick and Easy Decision making Does Not Get It Right, Proceedings from the 31st International Conference on Information Systems (ICIS), 2010.

[31] I. Blohm, U. Bretschneider, J. M. Leimeister, and H. Krcmar, Does Collaboration among

Participants Lead to Better Ideas in IT-Based Idea Competitions? An Empirical Investigation, 43rd Hawaii International Conference on System Sciences (HICCS), pp. 1-10, 2010.

[32] K. McCrimmon and C. Wagner, "Stimulating Ideas Through Creative Software, Management Science, vol. 40, no. 11, pp. 1514-1532, 1994.

[33] N. Franke and S. Schah, How communities support innovative activities: an exploration of assistance and sharing among end-users, Research Policy, vol. 32, no. 1, pp. 157-178, Jan. 2003.

[34] S. Besemer and K. O'Quinn, Analyzing Creative Products: Refinement and Test of a Judging Instrument, The Journal of Creative Behavior, vol. 20, no. 2, pp. 115-126, 1986.

[35] M. Nagasundaram and R. P. Bostrom, The structuring of creative processes using GSS: a framework for research, Journal of Management Information Systems, vol. 11, no. 3, pp. 87-114, 1994.

[36] D. L. Dean, J. M. Hender, T. L. Rodgers, and E. L. Santanen, Identifying quality, novel, and creative Ideas: Constructs and scales for idea evaluation, Journal of the Association for Information Systems, vol. 7, no. 10, pp. 646-699, 2006.

[37] Y.-H. Tseng, C.-J. Lin, and Y. I. Lin, Text mining techniques for patent analysis, Information Processing & Management, vol.43, no. 5, pp. 1216-1247, 2007.

[38] A. B. Clegg, Computational-Linguistic Approaches to Biological Text Mining, School of Crystallography Birkbeck, University of London, 2008.

[39] G. A. Wang, V. Tech, and W. Fan, A knowledge adoption model based framework for identifying helpful user-generated content in online communities, Proceedings from 32nd the International Conference on Information Systems (ICIS), pp. 1-11, 2011.

[40] R. Sharda and M. Henry, "Information Extraction from Interviews to Obtain Tacit Knowledge : A Text Mining Application Information Extraction from Interviews to Obtain Tacit knowledge, Proceedings from the 15th Americas Conference on Information Systems (AMCIS), pp. 1-12, 2009.

[41] M. Böhringer and P. Helmholz, What are they Thinking? - Accessing Collective Intelligence in Twitter, Proceedings from the 24th International Bled Conference, pp. 310-320, 2011.

[42] Q. Li, Y.-fang Brook, and Y.-fang B. Wu, Information Mining : Integrating Data Mining and Text Mining for Business Intelligence, Proceedings from the 12th Americas Conference on Information Systems (AMCIS), pp. 1410-1416, 2006.

[43] W. Cukier and E. J. Nesselroth-Woyzbun, Genres of Spam Genres of Spam Expectations and deceptions, Scandinavian Journal of Information Systems, vol. 20, no. 1, pp. 1-24, 2008.

[44] Y.-H. Tseng, Z.-P. Ho, K.-S. Yang, and C.-C. Chen, Mining term networks from text collections for crime investigation, Expert Systems with Applications, vol. 39, no. 11, pp. 10082-10090, 2012.

[45] 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.

[46] J. B. Lovins, Error Evaluation for Stemming Algorithms as Clustering Algorithms, Journal of the American Society for Information Science, vol. 22, no. 1, pp. 28-40, 1971.

[47] M. Steinbach, G. Karypis, and V. Kumar, A Comparison of Document Clustering Techniques, in Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.

[48] A. Hotho, A Brief Survey of Text Mining, Forum American Bar Association, pp. 19-62, 2005.

[49] T. Miller, Data and Text Mining: A Business Applications Approach. Upper Saddle River: Prentice-Hall, 2005, p. 455.

[50] C. Ding and X. He, K-means clustering via principal component analysis, Proceedings from 21st International Conference on Machine learning (ICML), pp. 1-9, 2004.

[51] F. Can and E. A. Ozkarahan, Concepts and Effectiveness of the Clustering Methodology for Text Databases, ACM Transaction on Database Systems, vol. 15, no. 4, pp. 483-517, 1990.

[52] G. Grefenstette, Explorations in Automatic Thesaurus Discovery. Norwell: Kluwer Academia, 1994.

[53] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Pearson Addison Wesley, 2006, p. 769.

[54] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search. Addison-Wesley Professional, 2010, p. 944.

[55] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Vol. 27, No. 8, pp. 861-874, Jun. 2006.

[56] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Boston: MIT Press, 1999, p. 680.

[57] A. Gelmann and J. Hill, Data Analysis Using Regression and Multilevel Models. Boston: Cambridge University Press, 2007, p. 625.

[58] C. M. Fuller, D. P. Biros, and D. Dursun, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection. Proceedings from the 41st Hawaii International Conference on Systems.