# Academic-Industrial Perspective on the Development and Deployment of a Moderation System for a Newspaper Website

**Dietmar Schabus, Marcin Skowron**
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, 1010 Vienna, Austria
{dietmar.schabus, marcin.skowron}@ofai.at

**Abstract**

This paper describes an approach and our experiences from the development, deployment and usability testing of a Natural Language Processing (NLP) and Information Retrieval system that supports the moderation of user comments on a large newspaper website. We highlight some of the differences between industry-oriented and academic research settings and their influence on the decisions made in the data collection and annotation processes, selection of document representation and machine learning methods. We report on classification results, where the problems to solve and the data to work with come from a commercial enterprise. In this context typical for NLP research, we discuss relevant industrial aspects. We believe that the challenges faced as well as the solutions proposed for addressing them can provide insights to others working in a similar setting. Data and experiment code related to this paper are available for download at https://ofai.github.io/million-post-corpus.

## 1. Background

For about two years, we have been working on an applied research project in a collaboration between a research institute and a large Austrian broadsheet newspaper (DER STANDARD), which supports the moderation of the comments posted to the newspaper's website[1] by its readers.

Like many newspaper websites, DER STANDARD's website features a comment section below each newspaper article, where users engage in discussion. In the year 2017, more than 9.5 million comments were posted by more than 55,000 distinct users. To ensure high quality in the discourse, the newspaper's community management department invests considerable effort in the moderation of the discussion fora, using both machine-learning-based tools and a team of professional human forum moderators.

With the project goal to improve the moderation, the moderators have defined eight relevant categories of posts, and have annotated a collection of posts with respect to these categories. The annotated categories are "negative sentiment", "positive sentiment", "off-topic", "inappropriate", "discriminating", "feedback", "personal stories" and "arguments used". The detailed description of the categories and the annotation process, as well as the resulting data set and the baseline classification results are provided in Schabus et al. (2017). Both the data set and the classification experiment code are available online for research purposes.[2]

We have designed, developed and deployed a moderator dashboard that provides various ways of searching, filtering, sorting and aggregating according to the introduced set of categories to help the moderators find locations in the discussions where moderation actions are required. In this process, we have addressed the following tasks that often interconnect predominately research- and industry-oriented aspects:

- Automatic labeling of new user comments according to the defined categories – a text classification problem,
- Using these predictions and other (meta-)data for finding posts and/or entire discussions that require moderator attention – an information retrieval problem,
- Providing a user interface to the moderators so they can use the results of the above in their workflow – user interface design,
- Integrating the resulting system into the existing IT infrastructure – system integration.

Since we want to deliver a (prototype) system that is usable in practice, our setting differs considerably from academic research in several aspects. For example, both user interface design and system integration are not typically relevant in NLP research.

## 2. Challenges

In this section, we describe a few challenges we have faced in detail and how the issues were addressed. We have grouped them under the four terms *holism*, *specificity*, *"messy" data* and *integration*, highlighting differences between academic and industrial settings.

### 2.1. Holism

In academic research we are often focused on a highly specific problem, and we can make extensive assumptions about aspects that are not in the immediate center of attention. In contrast, industrial endeavors require a more holistic view; they need to work with given practical settings and address specific requirements of live systems. In particular, we have identified three relevant perspectives to our project, all of which need to be considered simultaneously:

The *scientific/technical perspective* focuses on questions such as which methods to apply for particular subproblems, how to best make use of the available data, which evaluation metrics to apply, etc.

The *industrial perspective* focuses on the operational realization and deals with topics like integration and interfaces, software quality, performance and scalability, privacy, security, backups, etc.

[1] https://derstandard.at

[2] https://ofai.github.io/million-post-corpus

Finally, the *user perspective* is concerned with the benefits the system is able to deliver to the end-users, in our case the moderators working for the newspaper.
For example, using evaluation metrics like precision, recall and $F_1$-score for a classification problem is well-established in machine learning research (scientific perspective), but measuring the time savings for a well-defined moderation task tells us more adequately how well we are addressing the needs of the moderators (user perspective).
It has shown to be beneficial to frequently switch between these perspectives during the project timespan or to consider and address them simultaneously.

## 2.2. Specificity

Even though we have identified the requirement for holistic thinking as one challenge for our endeavor in the previous subsection, we can at the same time also name challenges that come from the highly specific practical needs in the given real-world setting. For example, the applied classification scheme (i.e., the annotated categories) could be criticized in a purely academic setting as being specifically tailored to the needs of one particular newspaper. Indeed, it is difficult to find related work that deals with text classification according to categories like "arguments used", "personal stories" or "feedback" originating from the concrete moderation needs at DER STANDARD. And even for our category negative/positive sentiment, the related research literature in sentiment analysis often relates to online *reviews* for different things like movies (Pang and Lee, 2005; Socher et al., 2013; Maas et al., 2011; Le and Mikolov, 2014), restaurants (Snyder and Barzilay, 2007) or books (Sakunkoo and Sakunkoo, 2009), where a clear indication of sentiment can be expected in every document, and the domain restriction can be expected to facilitate sentiment classification. In our data, sentiment has yet another specific meaning stemming from the application; in particular, the moderators are interested in locating negative sentiment in order to prevent escalation in user discussions.
In general, in a concrete industrial setting it is likely that one has to deal with very specific phenomena, use cases and goals, and results from academic research might not carry over directly.

## 2.3. "Messy" Data

Data annotation by humans is time-consuming and thus costly, especially when specific domain expertise is necessary, as is the case with the categories the moderators have defined with their particular moderation goals at DER STANDARD in mind. We need to ensure efficient use of moderator time in data annotation for model training and evaluation. Furthermore, most categories can be considered rare anomalies, resulting in strongly unbalanced data (e.g., the binary category "discriminating" has a prevalence of about 8% in our data set). These two factors explain the somewhat complicated, exploratory annotation procedure described in our data set paper (Schabus et al., 2017): In the first attempt, where 1,000 user comments were selected randomly for annotation, some categories were very weakly represented. Subsequently making use of the moderators' experience in selecting suitable topics (e.g., articles about the refugee crisis or gender mainstreaming for finding discriminating posts) turned out to be helpful for acquiring additional positive instances.[3] However, this also has unwanted side-effects; First, it means that many posts are annotated only according to one particular category, i.e., the data sets for the categories are mostly disjoint and consequently separate classification models must be trained, rather than a single multi-label model. Second, the class distributions in the labeled data are no longer necessarily indicative of the real class distributions in practice. And even if we accept these disadvantages, it still does not mean that we have vast amounts of data: In our data collection, even the better represented categories have less than 2,000 positive examples.
While in academic research settings methods are typically evaluated on carefully compiled benchmark data sets, which have reasonable balance and size, in practical industry applied research settings these might not always be available in a similar quality and quantity. Our situation is also different from what one might associate with an industrial setting typical for large companies which have less limitations in terms of available data or capacities to conduct large scale annotations. In the presented approach we thus focused on the efficient usage of the available data acknowledging its characteristics which are neither typical for academia nor for large enterprises.

## 2.4. Integration

The goal of the project is to deliver a prototype system applicable in practice, i.e., supporting the moderation of current online discussions. Therefore, a connection to the production forum system is required, such that the prototype works with the live stream of new postings in near real-time. To achieve this, the prototype needed to be integrated into the existing IT infrastructure at the newspaper.
The IT environments typically used in research institutions (operating systems, programming languages, software libraries, database systems, etc.) differ significantly from those used in commercial enterprises. In the former case, open-source libraries are often used, where new methodological advances become available quickly. In the latter case, systems from large commercial vendors are often preferred, with certifications, support services, etc. Letting researchers use the tools they are accustomed to is beneficial for flexibility and agility in experimental prototype development; on the other hand, a prototype using the same technologies as the existing environment facilitates integration. Our approach to this situation was to compromise: give the researchers flexibility in the core area of experimentation (e.g., machine learning frameworks), but adapt to the enterprise systems in other areas (e.g., database system).
Another important aspect of adding an experimental prototype from a collaborative research project to a production environment is (data) security and privacy. No enterprise would tolerate the risks involved with a prototype system directly manipulating its production databases for a service used by thousands of users every day. Therefore, the data

[3] Positive here means that the property in question is present, e.g., that a given posting does exhibit the characteristics of the "discriminating" category.

was mirrored to a dedicated database server for the prototype system, such that the production system is shielded from potential bugs. By restricting this mirroring to the data that is actually required, most privacy risks can be avoided (e.g., no personal data of users are mirrored).

In a practical setting, scalability and performance under peak loads become key factors. In our setting with up to 200 new comments per minute and eight different labels to predict, the run-time performance requirements for prediction (time and memory) influenced the choice of methods. By keeping the models for comment representation and classification small enough to all fit into main memory simultaneously, and by processing new comments in batches, we achieve a performance of almost 20,000 classified comments per minute on a machine with 16 cores.

Finally, we need to keep in mind that the system needs to be completed on time before the end of the project and operated and maintained by the industry partner after that. Therefore, clean code, suitable error handling and documentation are essential; these aspects typically can and are neglected in purely academic settings.

## 3. Experimental Results

To better illustrate some of the challenges we face in our concrete industrial setting, we report the results of new experiments using our data set, which extend the experiments from our previous work (Schabus et al., 2017). There, the most promising method was a (linear) support vector machine on a paragraph vector representation (Le and Mikolov, 2014), which we further investigate in our new experiments, using the old setup as a baseline.

The first extension we consider is to train two paragraph vector models (one using the distributed memory method and one using the distributed bag-of-words method) and to then represent each document by the concatenation of the two vectors, as proposed by Le and Mikolov (2014). We used a vector size of 100 dimensions for each of the two models instead of 300 as we did in the baseline setup as this turned out to be superior in preliminary experiments, and it also keeps the dimensionality from becoming too excessive when two vectors are concatenated.

Secondly, we add topic features to the representation: Each of our user comments belongs to a news article, and for each news article we have meta-data including a topic path such as *sports / motorsports / formula 1*. We have selected 17 top level topics (e.g., *sports*, *economy*, *science*, etc.) and added the resulting 17 binary dimensions indicating topic membership to the representation for each comment.

Finally, we compare support vector machines with linear kernels against Radial Basis Function (RBF) kernels, motivated by the hypothesis that the new composite feature space requires more complex decision boundaries for accurate classification.

The evaluation results of a 10-fold stratified cross-validation on the data set from (Schabus et al., 2017) are given in Table 1, where Method 1 represents the baseline results from our prior study. Methods 2–9 represent all combinations of the three configuration options described above. Note that Method 2 is identical to the baseline except with regard to the number of dimensions (100 vs. 300).

In terms of $F_1$-score, Method 9 (concatenation, topics and RBF kernel) outperforms the baseline on five of eight categories, and Method 8 (concatenation, topics and linear kernel) outperforms the baseline on an additional category ("negative sentiment"). For the two remaining categories "inappropriate" and "personal stories", the results of Method 9 are less than 0.01 below the baseline results.

Using the concatenated representation generally helps to improve the prediction results (Methods 4 and 5 vs. Methods 2 and 3), most noticeably for the categories "feedback" and "personal stories". Adding topic information also generally improves the prediction results (Methods 6 and 7 vs. Methods 2 and 3), this time most noticeably for categories "negative sentiment", "off-topic", "discriminating", "feedback" and "personal stories". Combining both representation extensions results in further improvement, especially for the "negative sentiment", "off-topic" and "feedback" categories. Even when Methods 8 and 9 are not the best performing, the differences are insignificant for practical settings and therefore we choose one of these two for deployment, favoring a more uniform overall setup.

With respect to the challenges discussed in Section 2., the specificity of the data we work with becomes apparent in the context of the experiments. For example, there are no directly applicable baseline results to compare against for most of our categories, with the exception of the two "sentiment" categories where extensive prior work exists. Here however the differences are in the definition and scope of the labels, hindering direct comparison of classification results. For example, Le and Mikolov (2014) report sentiment classification accuracies above 90% on a balanced data set of movie reviews, while our best result for negative sentiment in terms of minority class $F_1$-score (0.6063) corresponds to only 63% accuracy on our set of user comments, which are highly diverse in terms of topic, style, length, author intention, etc.

Finally, the integration aspect also plays a role in selecting the classification method to use in practice. For example, with deep LSTM models, which achieved competitive results in our previous work, we need to sequentially load separate large models onto a GPU for efficient classification, increasing the required efforts in operation and maintenance of the system after the "hand-over" to the industry partner. On the other hand, paragraph vectors are an efficient representation in our scenario, because they are computed only once for all eight categories, and then fed into separate SVM models. The resulting system is relatively light-weight and easier to deploy and maintain in the long term.

## 4. Conclusions

In this "industry track" paper, we have shared our experiences from a collaborative applied research project involving a small research institution and a medium size commercial enterprise. The goal of the project is to develop and deploy a prototype system that supports the moderation of user discussions on a large newspaper website. A key building block of this system is a text classification module predicting eight moderator-defined category labels. We have described a number of challenges faced in this context and

| | | Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Concat | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ |
| | | Topics | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |
| | | Kernel | Linear | RBF | Linear | RBF | Linear | RBF | Linear | RBF |
| Category | Measure | 1 (BL) | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Negative | Precision | 0.5842 | 0.5653 | 0.5755 | 0.5832 | 0.5893 | 0.5975 | 0.6106 | 0.6112 | **0.6216** |
| | Recall | 0.5624 | 0.5659 | 0.5228 | 0.5908 | 0.5192 | 0.5310 | 0.4684 | **0.6014** | 0.4837 |
| | $F_1$ | 0.5731 | 0.5656 | 0.5479 | 0.5870 | 0.5520 | 0.5623 | 0.5301 | **0.6063** | 0.5441 |
| Positive | Precision | 0.0397 | 0.0644 | 0.0707 | 0.0845 | 0.0618 | **0.1020** | 0.0851 | 0.0804 | 0.0977 |
| | Recall | **0.4651** | 0.3488 | 0.3023 | 0.2791 | 0.2558 | 0.3488 | 0.2791 | 0.2093 | 0.3023 |
| | $F_1$ | 0.0731 | 0.1087 | 0.1145 | 0.1297 | 0.0995 | **0.1579** | 0.1304 | 0.1161 | 0.1477 |
| OffTopic | Precision | 0.2065 | 0.1930 | 0.2039 | 0.2010 | 0.2090 | 0.2284 | **0.2579** | 0.2472 | 0.2524 |
| | Recall | **0.6241** | 0.5897 | 0.4552 | 0.5707 | 0.4724 | 0.5741 | 0.4759 | 0.6086 | 0.4534 |
| | $F_1$ | 0.3103 | 0.2908 | 0.2816 | 0.2973 | 0.2898 | 0.3268 | 0.3345 | **0.3516** | 0.3243 |
| Inappr | Precision | 0.1340 | 0.1074 | 0.1382 | 0.1218 | **0.1475** | 0.1203 | 0.1340 | 0.1179 | 0.1433 |
| | Recall | 0.5776 | 0.5347 | 0.4059 | 0.5116 | 0.4059 | **0.5974** | 0.4158 | 0.5248 | 0.4125 |
| | $F_1$ | **0.2175** | 0.1789 | 0.2062 | 0.1967 | 0.2164 | 0.2002 | 0.2027 | 0.1925 | 0.2128 |
| Discrim | Precision | 0.1111 | 0.1038 | 0.1206 | 0.1115 | 0.1402 | 0.1207 | 0.1343 | 0.1223 | **0.1547** |
| | Recall | 0.3936 | 0.4574 | 0.2057 | 0.4610 | 0.1844 | **0.5922** | 0.3440 | 0.5071 | 0.2837 |
| | $F_1$ | 0.1733 | 0.1692 | 0.1520 | 0.1796 | 0.1593 | **0.2005** | 0.1932 | 0.1971 | 0.2003 |
| Feedb | Precision | 0.5240 | 0.4604 | 0.5039 | 0.4865 | **0.5393** | 0.4520 | 0.4743 | 0.4839 | 0.5311 |
| | Recall | 0.7056 | 0.7233 | 0.6472 | 0.7317 | 0.7018 | **0.7679** | 0.7294 | 0.7633 | 0.7356 |
| | $F_1$ | 0.6014 | 0.5626 | 0.5666 | 0.5844 | 0.6099 | 0.5691 | 0.5748 | 0.5923 | **0.6168** |
| Personal | Precision | **0.6247** | 0.5525 | 0.5462 | 0.5995 | 0.5835 | 0.5563 | 0.5771 | 0.5952 | 0.5898 |
| | Recall | 0.8123 | 0.8160 | 0.8252 | 0.8271 | 0.8388 | 0.8394 | 0.8498 | 0.8332 | **0.8505** |
| | $F_1$ | **0.7063** | 0.6589 | 0.6574 | 0.6951 | 0.6882 | 0.6691 | 0.6874 | 0.6944 | 0.6966 |
| Argum | Precision | **0.5657** | 0.5636 | 0.5398 | 0.5594 | 0.5457 | 0.5631 | 0.5434 | 0.5581 | 0.5458 |
| | Recall | 0.6614 | 0.7114 | **0.7769** | 0.6722 | 0.7652 | 0.7250 | 0.7652 | 0.6908 | 0.7632 |
| | $F_1$ | 0.6098 | 0.6289 | 0.6370 | 0.6107 | **0.6371** | 0.6339 | 0.6355 | 0.6174 | 0.6365 |
| $>$ BL | Precision | | 1 | 3 | 2 | 6 | 4 | 5 | 4 | 6 |
| | Recall | | 5 | 2 | 5 | 2 | 5 | 3 | 5 | 3 |
| | $F_1$ | | 2 | 2 | 4 | 3 | 4 | 4 | 5 | 5 |

Table 1: Classification results: precision, recall and $F_1$-scores per method and category. BL indicates the baseline from Schabus et al. (2017). Values outperforming the baseline are underlined, the best value per row is in **bold**. The last three rows indicate the number of times the baseline was outperformed per method and measure.

grouped them under the terms holism, specificity, "messy" data and integration, highlighting identified differences between academic and industrial perspectives. Finally, we reported new results on our data set to illustrate some of these challenges and proposed solutions more concretely.

## 5. Acknowledgments

This research was partially funded by the Google Digital News Initiative.[4] We thank DER STANDARD and their moderators for the interesting collaboration.

## 6. Bibliographical References

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proc. ICML*, pages 1188–1196, Bejing, China.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proc. ACL*, pages 142–150, Portland, OR, USA.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. ACL*, pages 115–124, Ann Arbor, MI, USA.

Sakunkoo, P. and Sakunkoo, N. (2009). Analysis of social influence in online book reviews. In *Proc. AAAI*, pages 308–310, San Jose, CA, USA.

Schabus, D., Skowron, M., and Trapp, M. (2017). One million posts: A data set of German online discussions. In *Proc. SIGIR*, pages 1241–1244, Tokyo, Japan.

Snyder, B. and Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *Proc. ACL*, pages 300–307, Rochester, NY, USA.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*, pages 1631–1642, Seattle, WA, USA.

[4] https://www.digitalnewsinitiative.com