One Million Posts: A Data Set of German Online Discussions

Dietmar Schabus, Marcin Skowron, Martin Trapp
Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
firstname.lastname@ofai.at

ABSTRACT

In this paper we introduce a new data set consisting of user comments posted to the website of a German-language Austrian newspaper. Professional forum moderators have annotated 11,773 posts according to seven categories they considered crucial for the efficient moderation of online discussions in the context of news articles. In addition to this taxonomy and annotated posts, the data set contains one million unlabeled posts. Our experimental results using six methods establish a first baseline for predicting these categories. The data and our code are available for research purposes from https://ofai.github.io/million-post-corpus.

ACM Reference format:
Dietmar Schabus, Marcin Skowron, Martin Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan, ,* 4 pages.
DOI: 10.1145/3077136.3080711

1 INTRODUCTION

Publicly available large-scale data sets are a valuable asset for the research community, especially when they are annotated by human domain experts. We have recently completed assembling a large-scale data set consisting of one million user comments to news articles posted to the forum of the website of Der Standard, a large German-speaking newspaper. Professional forum moderators working for the newspaper have defined several relevant categories for the moderation of the online discussions and annotated 11,773 posts according to these categories.

While there are numerous data sets used for opinion mining and sentiment analysis, publicly available large-scale data sets that aim to capture other characteristics of user-generated content are relatively rare, especially for non-English textual online content. Examples of works that introduce different types of annotations than sentiment are frequently oriented towards the detection of abusive language. For example, Nobata et al. [10] describe a corpus of 2 million and 1.2 million posts randomly sampled from comments on Yahoo! Finance and News, and annotated by Yahoo! employees for abusive language. The annotation of abusive content was also the subject of the work presented by Waseem [14], where annotations of 6,909 tweets for hate speech were collected and compared. Bretschneider and Peters [2] presented a data set of 6,000 German Facebook posts annotated for abusive language towards six target groups (e.g., foreigners, politicians, the media).

In comparison to these prior works, the data set presented here introduces a wider set of categories deemed as crucial for the effective moderation of online discussions in a large-scale newspaper forum. These enable multifaceted analysis of online interactions including the discussion between users or the effects of the editorial staff and moderators' actions on the individual and group states. The presented data set also enables a range of practical applications, serving as a base for the development of dedicated classifiers and other tools supporting the effective moderation of real-world online discussion fora. Furthermore, it allows the evaluation of methods in information retrieval, natural language processing and related fields [1, 12, 15].

In the remainder of this paper we describe the new data set and report on the first results comparing state-of-the-art methods commonly applied to text classification tasks. Both the presented data set and code are available for research purposes.

2 BACKGROUND AND MOTIVATION

Der Standard is an Austrian daily broadsheet newspaper which had a circulation of more than 390,000 in the year 2015.[1] On the newspaper's website[2], there is a discussion section below each news article where readers engage in online discussions. In 2015 alone, 7.6 million posts were authored by more than 52,000 distinct users. One of the important goals for online communities is to ensure a high quality in the discourse. To this end, the newspaper's community management department invests considerable effort in moderation of the discussion fora, using both machine-learning-based tools and professional human forum moderators. In a current effort to improve the moderation, the moderators have defined seven relevant categories of posts (see Section 3 for details on the categories). In multiple rounds of annotations, they have labeled posts with respect to these categories, giving rise to classification and retrieval tasks: finding *individual posts* that fulfill certain criteria on one hand, and on the other hand finding *entire fora* (of which each corresponds to a news article) that overall show the most noticeable characteristics and need moderator attention.

In addition to the tasks described above, this real-world data set can support other information retrieval tasks, e.g., related to inter-post similarity, post-to-article similarity, user behavior patterns etc. The data set is furthermore useful to other fields including natural language processing, sentiment analysis and social network analysis. As most of the available data sets and textual resources are in English, the German data set presented here supports the advancement of tools and models for German language, and can also find its application as an additional benchmark data set in the evaluation of language-independent methods.

[1] According to http://www.media-analyse.at/table/2612 (2017-04-13)
[2] http://derstandard.at

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: 10.1145/3077136.3080711

Table 1: Annotations from all annotation rounds combined: Number of posts and percentages.

	Negative	Neutral	Positive	OffTopic	Inappr	Discrim	Feedb	Personal	Argum
Does apply	1691	1865	43	580	303	282	1301	1625	1022
Does not apply	1908	1734	3556	3019	3296	3317	4737	7711	2577
Total	3599	3599	3599	3599	3599	3599	6038	9336	3599
Percentage	47 %	52 %	1 %	16 %	8 %	8 %	22 %	17 %	28 %

3 DATA SET

Registered users can post comments below each news article on the newspaper website. Posts consist of a headline (max. 250 characters) and main body (max. 750 characters). The posts are publicly visible and display the username and timestamp in addition to the headline and body. Posts can be submitted as replies to earlier posts, giving rise to tree-like discussion thread structures. Furthermore, each post shows two counters, one for positive and one for negative votes by other community members. All of this information is included in the released corpus, with usernames anonymized to new numeric IDs. Forum moderators and also editorial staff of the newspaper participate in the discussions; the respective user IDs are indicated as newspaper staff in the data. Furthermore, we included the date of publication, headline, main body and topic of the news articles. When user posts violate the website's community guidelines (e.g., vulgar language), they either never appear publicly (pre-moderation) or they are taken offline at a later point in time (post-moderation). Our data set also contains such posts.

3.1 Annotated Categories

The newspaper editors and moderators have defined several goals for moderation. These goals define post categories of interest for automatic classification. The categories can be grouped into desirable, neutral and undesirable content, as follows.

Potentially undesirable content:

Sentiment An important goal is to detect changes in the prevalent sentiment in a discussion, e.g., the location within the fora and the point in time where a turn from positive/neutral sentiment to negative sentiment takes place.

Off-Topic Posts which digress too far from the topic of the corresponding article.

Inappropriate Swearwords, suggestive and obscene language, insults, threats etc.

Discriminating Racist, sexist, misogynistic, homophobic, antisemitic and other misanthropic content.

Neutral content that requires a reaction:

Feedback Sometimes users ask questions or give feedback to the author of the article or the newspaper in general, which may require a reply/reaction.

Potentially desirable content:

Personal Stories In certain fora, users are encouraged to share their personal stories, experiences, anecdotes etc. regarding the respective topic.

Arguments Used It is desirable for users to back their statements with rational argumentation, reasoning and sources.

Several of these categories are based on respective rules in the community guidelines of the website.[3] The sentiment category is ternary (positive/neutral/negative), all other categories are binary (yes/no).

3.2 Annotation Procedure

Four professional moderators participated in three rounds of annotations. The first round was meant as a dry-run for the annotation procedure; the results are not part of the final data corpus. In this round, three moderators annotated all 160 posts of a recent news article in parallel. The results were then discussed in order to clarify disagreements and to find a common definition of the categories.

The goals of the second annotation round were to determine the category distributions, to assess the inter-annotator agreement, and to obtain a first body of data. 1,000 posts were randomly selected from a 12 month period. Each of the three moderators annotated the 1,000 posts in parallel.

After discussing the results from the second annotation round, a third round was carried out, with the main goal of increasing the number of minority class samples. Therefore, posts were not selected randomly for the third annotation round, but in a way to increase the expected number of posts belonging to one of the categories, following three strategies:

Firstly, based on the available decision for each post whether it went online or was filtered out during pre-moderation, 10 news articles (2,599 posts) were selected for which the percentage of filtered-out posts was high, with the expectation of finding an increased number of posts from the following categories: negative sentiment, inappropriate, discriminating and off-topic.

Secondly, 13 discussions (5,737 posts) were selected from a special "share your thoughts" section of the forum, with the expectation of finding many posts that belong to the personal stories category.

Thirdly, 2,439 posts were selected for annotation that were either replied to by a newspaper staff member, or that were classified as feedback using a classifier trained on the data available so far, with the expectation of finding posts belonging to the feedback category.

Posts selected for annotation based on the first strategy were annotated considering all categories, the other posts were annotated considering the respective category of interest only. To maximize the amount of labeled data given the limited resources, the three annotators did not work in parallel in this round, but each of them got different posts to annotate.

Taking the majority vote per category for the posts from the second round and combining the result with the data from round three, there are 3,599 posts with annotations regarding all categories, an additional 5,737 posts with a decision regarding the personal stories

[3] http://derstandard.at/2934632/Forenregeln-Community-Richtlinien

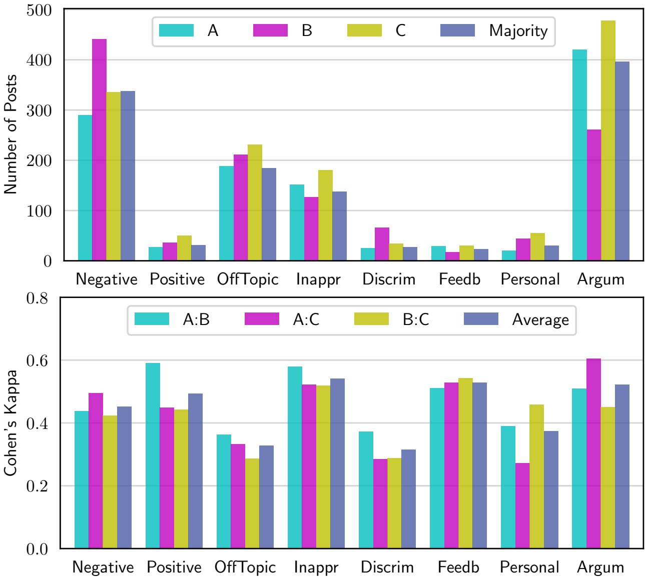

Figure 1: Results from the second annotation round (1,000 randomly selected posts). Category distribution (top) and inter-annotator agreement (bottom). Neutral sentiment category omitted.

category, and an additional 2,439 posts with a decision regarding the feedback category. The final distributions are given in Table 1. Counting every decision made by a moderator separately, the data set contains 58,568 expert judgments.

Considering only the 1,000 randomly selected posts of the second annotation round, we can estimate the distributions of the categories. Furthermore, we can compare the numbers of the three annotators, and compute Cohen's Kappa values to quantify the inter-annotator agreement. Figure 1 illustrates these points: we see that there are large differences in occurrence frequency between the categories, but these differences are generally quite consistent across the three annotators. The kappa values lie between 0.3 and 0.6, which qualifies as fair to moderate according to the guidelines by Landis and Koch [7]. Given the high complexity of the categories, we consider these inter-rater agreement values satisfactory.

3.3 Unlabeled Posts

Our data set contains one million unlabeled posts in addition to the 11,773 annotated by professional moderators. They facilitate the training of background models (e.g., word2vec [9] or doc2vec [8]) and semi-supervised modeling approaches [4]. Furthermore, they may be useful for additional tasks or analyses we have not considered so far.

For annotated posts that were individually selected for annotation, we want to include the context in the data set. Therefore, we include the other posts of the same discussion thread, which amounts to 38,747 unlabeled posts. For the remaining 961,253 to reach one million, entire fora were selected randomly for inclusion.

To conclude this section on the data set, Table 2 presents relevant numbers for the final corpus. Additional statistics will be available on the corpus website.

Table 2: Statistics of the data set.

Total number of posts	1,011,773
Number of unlabeled posts	1,000,000
Number of labeled posts	11,773
Number of category annotation decisions	58,568
Number of posts taken offline by moderators	62,320
Min/Median/Max post length (words)	0 / 21 / 500
Vocabulary size (≥ 5 occurrences)	129,070
Number of articles	12,087
Number of article topics	1,229
Number of users	31,413
Min/Median/Max number of posts per article	1 / 22 / 3,656
Min/Median/Max number of posts per topic	1 / 142 / 44,329
Min/Median/Max number of posts per user	1 / 5 / 4,682
Min/Median/Max number of users per article	1 / 15 / 1,371
Min/Median/Max number of users per topic	1 / 78 / 6,874
Number of pos./neg. community votes	3,824,806 / 1,096,300

4 EXPERIMENTS

In this section, we discuss our experiments with the new data set. We evaluate different methods for predicting the category labels, using two types of approaches: relying on the labeled data only, and also using the unlabeled posts.

4.1 Methods Using Labeled Data Only

A typical baseline approach for this type of problem is to use a Bag Of Words (BOW) representation and a linear support vector machine [5] for classification.

The probabilistic method of Multinomial Naive Bayes (MNB) is another classic choice for text categorization. Wang and Manning [13] showed in 2012 that BOW and MNB were still competitive classifiers for this kind of task. Furthermore, they proposed a new variant of using naive Bayes log-count ratios as features for a support vector machine, a combination of the two methods that we also include (NBSVM).

4.2 Methods Using Labeled and Unlabeled Data

In an attempt to both reduce the representation dimensionality and to make use of the large number of unlabeled posts, we have computed a word2vec embedding [9] on the unlabeled posts. As an intermediate step, this results in an embedding of the 129,070 distinct words from the large corpus in a 300-dimensional space, where semantically similar words should end up as close lying points. Next we cluster the points in this space into 1,000 clusters using the K-means algorithm. Finally, we replace each word in the labeled posts with the ID of the cluster the word belongs to, and compute a bag of cluster ID (BOCID) representation for all labeled posts, similar to [3, 11]. Finally we train a support vector machine on this representation. This way, the dimensionality of the representation is reduced from 129,070 (BOW) down to 1,000 (BOCID).

With similar motivation, we have trained a doc2vec (D2V) document embedding [8] on the unlabeled posts. This results in a mapping function that allows us to compute 300-dimensional real-valued representations for entire documents (i.e., posts). Then, this

Table 3: Classification results: precision, recall and F_1-scores per method and category. Best value per row in bold.

Category	Meas.	BOW	MNB	NBSVM	BOCID	D2V	LSTM
Negative	Prec.	0.5521	0.5637	0.5660	0.5345	**0.5842**	0.5349
	Rec.	0.5109	0.4867	0.4512	0.5452	0.5624	**0.7197**
	F_1.	0.5307	0.5224	0.5021	0.5398	0.5731	**0.6137**
Positive	Prec.	0.1000	0.0000	**0.2353**	0.0662	0.0397	0.0000
	Rec.	0.0698	0.0000	0.0930	0.2093	**0.4651**	0.0000
	F_1.	0.0822	0.0000	**0.1333**	0.1006	0.0731	0.0000
OffTopic	Prec.	0.2754	**0.6190**	0.3969	0.2252	0.2065	0.2742
	Rec.	0.2379	0.0224	0.1328	0.5121	**0.6241**	0.2638
	F_1.	0.2553	0.0433	0.1990	**0.3128**	0.3103	0.2689
Inappr	Prec.	0.1627	0.0000	0.1765	0.1516	0.1340	**0.1964**
	Rec.	0.1122	0.0000	0.0495	0.3993	**0.5776**	0.1089
	F_1.	0.1328	0.0000	0.0773	**0.2198**	0.2175	0.1401
Discrim	Prec.	0.1847	0.0000	**0.2683**	0.1301	0.1111	0.1136
	Rec.	0.1028	0.0000	0.0780	0.2943	**0.3936**	0.1418
	F_1.	0.1321	0.0000	0.1209	**0.1804**	0.1733	0.1262
Feedb	Prec.	0.6554	**0.7465**	0.7356	0.5094	0.5240	0.6307
	Rec.	0.5803	0.4074	0.5219	0.6879	**0.7056**	0.6287
	F_1.	0.6156	0.5271	0.6106	0.5853	0.6014	**0.6297**
Personal	Prec.	**0.6981**	0.5491	0.6916	0.5762	0.6247	0.6380
	Rec.	0.5920	0.4578	0.4788	0.7120	**0.8123**	0.6658
	F_1.	0.6407	0.4993	0.5658	0.6369	**0.7063**	0.6516
Argum	Prec.	**0.6105**	0.5086	0.6064	0.5642	0.5657	0.5685
	Rec.	0.5215	0.3170	0.4628	0.6106	**0.6614**	0.6458
	F_1.	0.5625	0.3906	0.5250	0.5865	**0.6098**	0.6047
Wins	Prec.	2	2	2	0	1	1
	Rec.	0	0	0	0	7	1
	F_1	0	0	1	3	2	2

mapping function is applied to the labeled posts and a support vector machine is trained on this data.

Finally, we evaluate a neural network architecture using a Long Short-Term memory (LSTM) layer [6]. We use a pre-trained embedding layer with the results of the word embedding on the unlabeled posts. Next we use an LSTM layer with 128 units, followed by a fully connected layer and a softmax layer to predict the applicability of the category in question.

4.3 Results

To compare the methods and provide a baseline for our data set, we have carried out a stratified 10-fold cross validation experiment using the six methods described above. We report classification results using precision, recall and F_1-score on the minority class as performance measures. Table 3 shows the values per category and method. We see that overall, the methods using both labeled and unlabeled data (BOCID, D2V and LSTM) achieve higher values for recall and also F_1. The weakly represented category positive sentiment is the only category where the best method (considering F_1) uses labeled data only. D2V is the best method overall, especially if we consider recall more important than precision. As expected, the number of minority class instances has a significant influence on the classification accuracy (compare Table 1).

5 SUMMARY AND CONCLUSIONS

We have presented a new data set consisting of user comments posted to the website of Der Standard, an Austrian newspaper. 11,773 posts have been annotated by professional forum moderators according to seven categories relevant for the moderation of online discussions. The data set furthermore includes one million unlabeled posts. In addition to the textual content of the post and the annotations, the data set includes several other aspects like metadata (user ID, timestamp, thread structure, news article) and community reactions (negative/positive votes per post). Regarding the categories, we evaluated six methods to predict the applicability of each category given the post text. Methods using both labeled and unlabeled data outperform methods using labeled data only. Despite moderate inter-annotator agreement, we achieved satisfactory results which can serve as a baseline.

Future work will include additional annotations as well as benchmarking semi-supervised methods and additional neural network architectures on this data set.

ACKNOWLEDGMENTS

This research was partially funded by the Google Digital News Initiative (https://www.digitalnewsinitiative.com). We thank Der Standard and their moderators for the interesting collaboration and the annotation of the presented corpus. The GPU used for this research was donated by nvidia.

REFERENCES

[1] Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2016. Identifying the role of individual user messages in an online discussion and its use in thread retrieval. *Journal of the Association for Information Science and Technology* 67, 2 (2016), 276–288.

[2] Uwe Bretschneider and Ralf Peters. 2017. Detecting Offensive Statements towards Foreigners in Social Media. In *Proceedings of HICSS*. Waikoloa Village, HI, USA, 2213–2222.

[3] Alexander M. Bronstein, Michael M. Bronstein, Leonidas J. Guibas, and Maks Ovsjanikov. 2011. Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval. *ACM Trans. Graph.* 30, 1 (2011), 1–20.

[4] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. 2010. *Semi-Supervised Learning* (1st ed.). The MIT Press.

[5] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[7] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[8] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of ICML*. Beijing, China, 1188–1196.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*. Scottsdale, AZ, USA.

[10] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of WWW*. Montreal, Canada, 145–153.

[11] Josef Sivic and Andrew Zisserman. 2003. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of ICCV*. Nice, France, 1470–1477 vol.2.

[12] Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011. Learning online discussion structures by conditional random fields. In *Proceedings of SIGIR*. Beijing, China, 435–444.

[13] Sida Wang and Christopher D. Manning. 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In *Proceedings of ACL*. Stroudsburg, PA, USA, 90–94.

[14] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of EMNLP Workshop on NLP and Computational Social Science*. Austin, Texas, USA, 138–142.

[15] Tim Weninger. 2014. An exploration of submissions and discussions in social news: Mining collective intelligence of Reddit. *Social Network Analysis and Mining* 4, 1 (2014), 1–19.