Prediction for the Newsroom: Which Articles Will Get the Most Comments?

Carl Ambroselli[1], Julian Risch[1], Ralf Krestel[1], and Andreas Loos[2]

[1]Hasso-Plattner-Institut, University of Potsdam, Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Germany
[1]carl.ambroselli@student.hpi.de, julian.risch@hpi.de, ralf.krestel@hpi.de
[2]ZEIT online, Askanischer Platz 1, 10963 Berlin, Germany
[2]andreas.loos@zeit.de

Abstract

The overwhelming success of the Web and mobile technologies has enabled millions to share their opinions publicly at any time. But the same success also endangers this freedom of speech due to closing down of participatory sites misused by individuals or interest groups. We propose to support manual moderation by proactively drawing the attention of our moderators to article discussions that most likely need their intervention. To this end, we predict which articles will receive a high number of comments. In contrast to existing work, we enrich the article with metadata, extract semantic and linguistic features, and exploit annotated data from a foreign language corpus. Our logistic regression model improves F1-scores by over 80% in comparison to state-of-the-art approaches.

1 Exploding Comment Threads

In the last decades, media and news business underwent a fundamental shift, from one-directional to bi-directional communication between users on the one side and journalists on the other. The use of social media, blogs, and the possibility to immediately share, like, and comment digital content transformed readers into active and powerful agents in the media business. This shift from passive "consumers" to active "agents" deeply impacts both media and communication science and has many positive aspects.

However, the possibilities and powers can also be misused. Pressure groups, lobbyists, trolls, and others are effectively trying to influence discussions according to their (very different) interests. An easy approach consists in burying unwanted arguments or simply destroying a discussion by blowing it up. After such an attack, readers have to crawl through hundreds of nonsense and meaningless comments to extract meaningful and interesting arguments. Blowing up a thread can be achieved by injecting provocative (but not necessarily off-topic) arguments into discussions. Bystanders are completing the goal of the destroyers, and they do so often unknowingly: with each — often well-intentioned — reaction to the provocation, they make it more difficult for others to follow the actual argumentation path and/or tree.

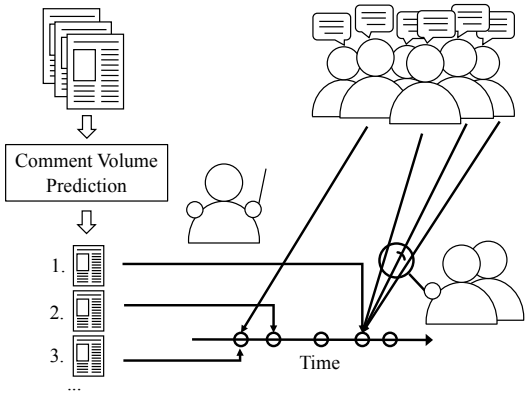

Figure 1: Integration of comment volume prediction into the newsroom workflow.

It is costly in terms of working power and time to keep the discussion area of a news site clean from attacks like that, and to watch the compliance of users ("netiquette"). As a reaction, many large online media sites worldwide closed their discussion areas or downsized them significantly (prominent examples of the last years are the Internet Movie Database, Bloomberg or the US-American National Public Radio). Other news provider and media sites, including us, take a different approach: A team of editors reads and filters comments on a 24/7-basis. This results in a huge workload with several thousand reader comments published each day. In its lifetime, an article receives between less than ten and more than 1500 comments; typical are about 100 to 150 comments. The number of published comments

presumably depends to a large extent on time, weather, and season as well as for each article on subject, length, style of writing, and author, among others.

Being able to predict which articles will receive high comment volume would be beneficial at two positions in the newsroom:

1. for the news director to schedule the publication of news stories, and

2. for scheduling team sizes and guiding the focus of the comment moderators and editors.

Figure 1 gives an overview of how comment volume prediction can be integrated into the workflow of a modern online news site. The incoming news articles are ranked based on the estimated number of comments they will attract. The news director takes these numbers into account in the decision process when to schedule which article for publication. This can balance the distribution of highly controversial topics across a day, giving not only readers and commenters the possibility to engage in each single one, but also distribute the moderation workload for comment editors evenly. Further, knowing which articles will receive many comments can help in the moderation process. Guiding the main focus of attention of moderators towards controversial topics not only facilitates efficient moderation, but also improves the quality of a comment thread. Our experience has shown that moderators entering the online discussion at an early stage can help keeping the discussion focused and fruitful.

In this paper, we study the task of identifying the weekly top 10% articles with the highest comment volume. We consider a new real-world dataset of 7 million news comments collected over more than nine years. In order to enrich our dataset and increase its meaningfulness, we propose to transfer a classifier trained on the English-language Yahoo News Annotated Comments Corpus (Napoles et al., 2017b) to our German-language dataset and leverage the additional class labels for comments in a post-publication prediction scenario. Experiments show that our logistic regression model based on article metadata, linguistic, and topical features outperforms state-of-the-art approaches significantly. Our contributions are summarized as (1) a transfer learning approach to learn early comments' characteristics, (2) an analysis of a new 7-million-comment dataset and (3) an improvement of F1-score by 81% compared to state-of-the-art in predicting most commented articles.

2 Related Work

Related work on newsroom assistants focuses on comment volume prediction for pre-publication and post-publication scenarios. By the nature of news articles, the attention span after article publication is short and in practice post-publication prediction is valuable only within a short time frame. Tsagkias et al. (2009) classify online newspaper articles using random forests. First, they classify whether an article will receive any comments at all. Second, they classify articles as receiving a high or low amount of comments. The authors find that the second task is much harder and that predicting the actual number of comments is practically infeasible. Badari et al. (2012) conclude the same, analyzing Twitter activity as a popularity indicator for news: Predicting popularity as a regression task results in large errors. Therefore, the authors predict classes of popularity by binning the absolute numbers (1-20, 20-100, 100-2400 received tweets). However, predicting the number of received tweets includes modeling both, the user behavior and the platform, which is problematic. It is part of a platform's business secrets how content is internally ranked and distributed to users, making it hard to distinguish cause and effect from the outside. In our scenario, we even see no benefit in predicting the exact number of comments. Instead, we predict which articles belong to the weekly top 10% articles with the highest comment volume, which is one of the tasks defined by Tsagkias et al. (2009).

In a post-publication scenario, Tsagkias et al. (2010) consider the comments received within the first ten hours after article publication. Based on this feature, they propose a linear model to predict the final number of comments. Comparing comment behavior at eight online news platforms, they observe seasonal trends. Tatar et al. (2011) consider the shorter time frame of five hours after article publication to predict article popularity. They also use a linear model and find that neither adding publication time and article category to the feature set nor extending the dataset from three months to two years improves prediction results. Their survey on popularity prediction for web content summarizes features with good predictive capabilities

and lists fields of application for popularity prediction (Tatar et al., 2012).

Rizos et al. (2016) focus on user comments to predict a discussion's controversiality. They extract a comment tree and a user graph from the discussion and investigate for example comment count, number of users, and vote score. The demonstrated improvement of popularity prediction with this limited, focused features motivates us to further explore content-based features of comments in our work.

Recently, research on deep learning (Nobata et al., 2016; Pavlopoulos et al., 2017) addresses (semi-) automation of the entire moderation task, but we see several issues that prevent us from putting these approaches into practice. First, the accuracy of these methods is not high enough. For example, reported recall (0.79) and precision (0.77) at the task of abusive language detection (Nobata et al., 2016) are not sufficient for use in production. With this recall, an algorithm would let pass every fifth inappropriate comment (containing hate speech, derogatory statements, or profanity), which is not acceptable. Pavlopoulos et al. (2017) address this problem by letting human moderators review comments that an algorithm could not classify with high confidence. Second, acceptance of these kind of black-box solutions is still limited in the community and the models lack comprehensibility. A compromise can be (ensemble) decision trees, because they achieve comparable results and can give reasons for their decisions (Kennedy et al., 2017). Still, moderators and users do not feel comfortable with machines deciding which comments are allowed to be published – not least because of fear of concealed censorship or bias.

3 Predicting High Comment Volume

For each news article, we want to predict whether it belongs to the weekly top 10% articles with the highest comment volume. We chose this relative amount to account for seasonal fluctuations and also to even out periods with low news worthiness. This traditional classification setting enables us to use established methods, such as logistic regression, to solve the task and provide explanations on why a particular article will receive many comments or not.

As a baseline to compare against, we implemented a random forest model with features from

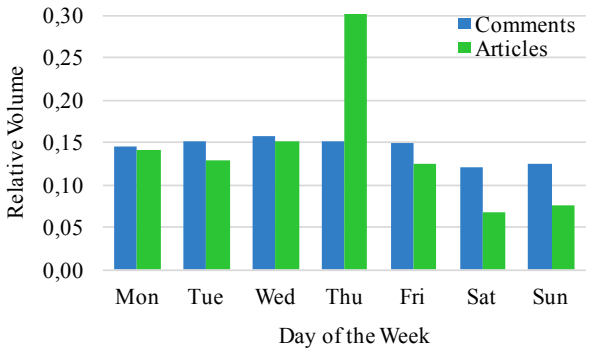

Figure 2: The number of received comments is not affected by a peek of article publications on Thursdays.

Tsagkias et al. (2009). For our approach we extend this feature set and categorize the features into five groups. Our **metadata** features consist of article publication time, day of the week, and whether the article is promoted on our Facebook page. We consider temperature and humidity during the hour of publication[1] and the number of "competing articles" as **context** features. Competing articles is the number of similar articles and the total number of articles published by our newspaper in the same hour. These articles compete for readers and user comments. Figure 2 visualizes how the number of received comments is not affected by the significantly higher number of published articles on Thursdays. The publication peek on Thursdays is caused by articles that are published in our weekly printed edition and at the same time published online one-to-one. Further, we incorporate **publisher** information, such as genre, department, and which news agency served as a source for the article. We include these features in order to study their impact and performance at comment volume prediction tasks and not in order to focus on engineering complex features.

In addition, we propose to leverage the article content itself. Starting with **headline** features, we use ngrams of length one to three as well as author provided keywords for the article. To capture topical information in the **body**, we rely on topic modeling and document embedding besides traditional bag-of-word (BOW) features. These guarantee that we also grasp some semantic representations of the articles. To this end, topic distributions, document embeddings, and word n-grams serve as semantic representa-

[1] as obtained for three large German cities, Berlin, Hamburg, and Frankfurt from http://www.dwd.de/

Table 1: Precision (P), recall (R), and F1-score for prediction of weekly top articles on the validation set.

Features	P	R	F1
Metadata	.12	.72	.21
Publication Time	.12	.74	.21
Promoted on Facebook	.29	.02	.01
Context	.13	.59	.22
Competing Articles	.11	.94	.20
Temperature and Humidity	.12	.27	.17
Publisher	.17	.85	.28
Author	.11	.96	.19
Genre	.16	.17	.17
Department	.15	.91	.26
Sources	.10	.38	.16
Medium	.11	.86	.20
Editor	.12	.82	.21
Headline	.15	.99	.26
Ngram 1-3 Words	.23	.48	.31
Keywords	.21	.57	.30
Body			
Doc2vec	.17	.63	.27
Stemmed BOW	.27	.61	.38
Topic model	.20	.66	.30

tions of articles. In order to model topics of news article bodies, we apply standard latent Dirichlet allocation (Blei et al., 2003). For the document embedding, we use a Doc2Vec implementation that downsamples higher-frequency words for the composition (Mikolov et al., 2013). We choose the vector length, number of topics, and window size based on F1-score evaluation on a validation set.

Despite recent advances of deep neural networks for natural language processing, there is a reason to focus on other models: For the application in newsrooms and the integration in semi-automatic processes, comprehensibility of the prediction results is very important. A black-box model — even if it achieved better performance — is not helpful in this scenario. Human moderators need to understand *why* the number of comments is predicted to be high or low. This comprehensibility issue justifies the application of decision trees and regression models, which allow to trace back predictions to their decisive factors. Table 1 lists precision, recall, and F1-score for the prediction of weekly top 10% articles with the highest comment volume. Especially the bag-of-words (BOW) and the topics of the article body, but also headline keywords and publisher metadata achieve higher F1-score than the metadata features. The highest precision is achieved with the binary feature whether an article is promoted on Facebook, whereas author and competing articles achieve the highest recall.

3.1 Automatic Translation of Comments

Whether the first comment is a provocative question in disagreement with the article or an off-topic statement influences the route of further conversation. We assume that this assumption holds not only for social networks (Berry and Taylor, 2017), but also for comment sections at news websites. Therefore, we consider the tone and sentiment of the first comments received shortly after article publication as an additional feature. Typical layouts of news websites (including ours) list comments in chronological order and show only the first few comments to readers below an article. Pagination hides later received comments and most users do not click through dozens of pages to read through all comments. As a consequence, early comments attract a lot more attention and, with their tone and sentiment, influence comment volume to a larger extent. Presumably, articles that receive controversial comments in the first few minutes after publication are more likely to receive a high number of comments in total.

To classify comments as controversial or engaging, we need to train a supervised classification algorithm, which takes thousands of annotated comments. Such training corpora exist, if at all, mostly for English comments, while our comments are written in German. We propose to apply machine translation to overcome this language barrier: Given a German comment, we automatically translate it into English. From a classifier that has been trained on an annotated English dataset, we can derive automatic annotations for the translated comment. The derived annotations serve as another feature for our actual task of comment volume prediction.

We reimplemented the classifier by Napoles et al. (2017a) and train on their English dataset. The considered annotations consist of 12 binary labels: addressed audience (reply to a particular user or broadcast message to a general audience), agreement/disagreement with previous comment, informative, mean, controversial, persuasive, off-topic regarding the corresponding news article, neutral, positive, negative, and mixed sentiment. We au-

Table 2: ZOCC is of similar structure as YNACC but contains 700 times more labeled comments.

	YNACC	ZOCC
Comments	9160	6,831,741
Comment Threads	2400	192,647

tomatically translate all comments in our German dataset into English using the DeepL translation service[2]. For the translated comments, we automatically generate annotations based on Napoles et al.'s classifier. Thereby, we transfer the knowledge that the classifier learned on English training data to our German dataset despite its different language. This approach builds on the similar content style of both corpora, which is described in the next section.

4 Dataset

We consider two datasets that both contain user comments received by news articles with similar topics. First, our German 7-million-comment dataset, which we call Zeit Online Comment Corpus (ZOCC)[3] and second, the English 10k-comment Yahoo News Annotated Comments Corpus (YNACC) (Napoles et al., 2017b). ZOCC consists of roughly 200,000 online news articles published between 2008 and 2017 and 7 million associated user comments in German. Out of 174,699 users in total, 60% posted more than one comment, 23% more than 10 comments and 7% more than 100 comments. For both, articles and comments, extensive metadata is available, such as author list, department, publication date, and tags (for articles) and user name, parent comment (if posted in response), and number of recommendations by other users (for comments). Not surprisingly, ZOCC is following a popularity growth with an increasing number of articles and comments over time. While our newspaper published roughly 1,300 articles per month in 2010 and each article received roughly 20 comments on average, we nowadays publish roughly 1,500 articles per month, each receiving 110 comments on average. As both corpora's articles and comments cover a similar time span of several years and many different departments, they deal with a broad range of topics. While the majority of articles in YNACC is about economy, ZOCC's major department is politics. More than 50% of the comments in ZOCC are posted in response to articles in the politics department, whereas in YNACC culture, society, and economy share an almost equal amount of around 20% each and politics on forth rank with 12%. On average, an article in ZOCC receives 90% of its comments within 48 hours, while it takes 61 hours for an article in YNACC. Despite their slight differences, both corpora cover most popular departments, which motivates the idea to transfer a classifier trained on YNACC to ZOCC. For YNACC, Napoles et al. propose a machine learning approach to automatically identify engaging, respectful, and informative conversations (2017a). By identifying weekly top 10% articles with the highest comment volume, we focus on a different task. Nonetheless, both corpora, ZOCC and YNACC, have similar properties: both corpora contain user comments posted in reaction to news articles across similar time span and similar topics. However, only the much smaller YNACC provides detailed annotations regarding, for example, comments' tone and sentiment.

5 Evaluation

We compare to the approach by Tsagkias et al. and evaluate on the same task (Tsagkias et al., 2009, 2010). Therefore, we consider a binary classification task, which is to identify the weekly top 10% articles with the largest comment volume. Table 3 lists our final evaluation results on the hold-out test set. We choose F1-score as our evaluation metric, since precision and recall are equally relevant in our scenario. On the one hand, we want to achieve high recall so that no important article and its discussion is overlooked. On the other hand, we have limited resources and cannot afford to moderate each and every discussion. A high precision is crucial so that our moderators focus only on articles that need their attention. All experiments are conducted using time-wise split with years 2014 to 2016 for training, January 2017 to March 2017 for validation, and April 2017 for testing. We find that our additional article and metadata features, but also the automatically annotated first comments outperform the baseline. Due to the diversity of the different features, their combination further improves the prediction results. In comparison to the approach by Tsagkias et al., we finally achieve an 81% larger F1-score.

[2] https://deepl.com
[3] http://www.zeit.de/

Table 3: Precision (P), recall (R), and F1-score of the baseline, all article and metadata features, annotations of comments shown on the first page, and all combined.

Features	P	R	F1
Tsagkias et al.	0.16	0.72	0.26
Article and metadata	0.26	0.75	0.39
1st page comments	0.29	0.50	0.36
Combined approach	0.42	0.52	0.47

Table 4: Precision (P) and recall (R) decline slightly after translation from English (E) into German (G).

Label	P(E)	R(E)	P(G)	R(G)
audience	.80	.80	.81	.82
agreement	.76	.18	.65	.09
informative	.55	.71	.51	.85
mean	.64	.52	.52	.37
controversial	.61	.90	.58	.94
disagreement	.60	.75	.58	.81
persuasive	.51	.89	.44	.97
off_topic	.67	.57	.66	.40
neutral	.68	.35	.62	.41
positive	.46	.13	.80	.10
negative	.70	.93	.71	.92
mixed	.45	.52	.40	.78

5.1 Automatically Translated Comments

With another experiment, we study the classification error introduced by translation. Therefore, we train two classifiers with the approach by Napoles et al.: First, we train and test a classifier on the original, English YNACC. Second, we automatically translate all comments in YNACC from English into German and use this translated data for training and testing of the second classifier. Comparing these two classifiers, we find that both precision and recall slightly decrease after translation, as shown in Table 4. Based on this result, we can assume that the translation of German comments into English introduces only a small error. Although YNACC and ZOCC differ in language, we can transfer a classifier that has been trained on YNACC to ZOCC. For each article, we use the labels assigned to the first four comments, which are visible on the first comment page below an article. The first four comments are typically received within very few minutes after article publication.

Table 5: Prediction of weekly top articles based on the number of comments received in the first x minutes after article publication.

Number of received comments	F1
after 2min	0.03
after 4min	0.03
after 8min	0.17
after 16min	0.33
after 32min	0.41
after 64min	0.45
sequence (after 2, 4, 8, 16, 32, 64min)	0.46

5.2 Number of Early Comments

As a baseline feature for comparison, we use the number of comments[4] received in a short time span after article publication. Annotated first page comments, but also article and metadata features significantly outperform the baseline until 32 minutes after article publication. After 32 minutes, the number of received comments outperforms every single feature (but not the combination of all our features). This is because the difference between final number of comments and so far received comments converges over time.

6 Conclusions

In this paper, we studied the task of predicting the weekly top 10% articles with the highest comment volume. This prediction helps to schedule the publication of news stories and supports moderation teams in focusing on article discussions that require most likely their attention. Our supervised classification approach is based on a combination of metadata and content-based features, such as article body and topics. Further, we automatically translate German comments into English to make use of a classifier pre-trained on English data: We classify the tone and sentiment of comments received in the first minutes after article publication, which improves prediction even further. On a 7-million-comment real-world dataset our approach outperforms the current state-of-the-art by over 81% larger F1-score. We hope that our prediction will help to reduce the number of cases where newspapers have no other choice but to close down a discussion section because of limited moderation resources.

[4]To allow for non-linear correlations, we pass the number of comments as an absolute count and a squared count.

References

Roja Bandari, Sitaram Asur, and Bernardo Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. volume 12, pages 26–33.

George Berry and Sean J Taylor. 2017. Discussion quality diffuses in the digital public square. In *Proceedings of the International Conference on World Wide Web (WWW)*. ACM, pages 1371–1380. https://doi.org/10.1145/3038912.3052666.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the Workshop on Abusive Language Online*. Association for Computational Linguistics, pages 73–77. http://www.aclweb.org/anthology/W17-3011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. Curran Associates Inc., USA, pages 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

Courtney Napoles, Aasish Pappu, and Joel R. Tetreault. 2017a. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. pages 628–631.

Courtney Napoles, Joel R. Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017b. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW@EACL)*. pages 13–23. https://doi.org/10.18653/v1/W17-0802.

Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the International Conference on World Wide Web (WWW)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pages 145–153. https://doi.org/10.1145/2872427.2883062.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the Workshop on Abusive Language Online*. Association for Computational Linguistics, pages 25–35. http://www.aclweb.org/anthology/W17-3004.

Georgios Rizos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2016. Predicting news popularity by mining online discussions. In *Proceedings of the International Conference on World Wide Web (WWW)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pages 737–742. https://doi.org/10.1145/2872518.2890096.

Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida. 2012. Ranking news articles based on popularity prediction. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE Computer Society, Washington, DC, USA, pages 106–110. https://doi.org/10.1109/ASONAM.2012.28.

Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. 2011. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS)*. ACM, New York, NY, USA, pages 67:1–67:8. https://doi.org/10.1145/1988688.1988766.

Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. 2009. Predicting the volume of comments on online news stories. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*. ACM, New York, NY, USA, pages 1765–1768. https://doi.org/10.1145/1645953.1646225.

Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. 2010. News comments: Exploring, modeling, and online prediction. In *Proceedings of the European Conference on Information Retrieval Research (ECIR)*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 191–203. https://doi.org/10.1007/978-3-642-12275-0_19.