A Dataset of Journalists' Interactions with Their Readership: When Should Article Authors Reply to Reader Comments?

Julian Risch
julian.risch@hpi.de
Hasso Plattner Institute, University of Potsdam
Germany

Ralf Krestel
ralf.krestel@hpi.de
Hasso Plattner Institute, University of Potsdam
Germany

ABSTRACT

The comment sections of online news platforms are an important space to indulge in political conversations andto discuss opinions. Although primarily meant as forums where readers discuss amongst each other, they can also spark a dialog with the journalists who authored the article. A small but important fraction of comments address the journalists directly, e.g., with questions, recommendations for future topics, thanks and appreciation, or article corrections. However, the sheer number of comments makes it infeasible for journalists to follow discussions around their articles in extenso. A better understanding of this data could support journalists in gaining insights into their audience and fostering engaging and respectful discussions. To this end, we present a dataset of dialogs in which journalists of The Guardian replied to reader comments and identify the reasons why. Based on this data, we formulate the novel task of recommending reader comments to journalists that are worth reading or replying to, i.e., ranking comments in such a way that the top comments are most likely to require the journalists' reaction. As a baseline, we trained a neural network model with the help of a pair-wise comment ranking task. Our experiment reveals the challenges of this task and we outline promising paths for future work. The data and our code are available for research purposes from: https://hpi.de/naumann/projects/repeatability/text-mining.html.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; • **Computing methodologies** → *Language resources*; Discourse, dialogue and pragmatics; Supervised learning; • **Human-centered computing** → **Social media**.

KEYWORDS

online discussions, news, text mining, natural language processing

ACM Reference Format:
Julian Risch and Ralf Krestel. 2020. A Dataset of Journalists' Interactions with Their Readership: When Should Article Authors Reply to Reader Comments?. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412764

1 COMMENTS ON NEWS PLATFORMS

Not long ago the interaction between newspapers and their readership was mostly unidirectional. A reader's letter to the editor was a time-consuming task and therefore a rare exception. The editor could decide to publish the letter in the next issue of the newspaper together with a statement or reply.

Nowadays, online news platforms offer comment sections, where any reader can easily post comments and discuss article topics at any time from anywhere. Some comments on these platforms directly address the journalist who wrote the article. Once posted, readers expect the journalists to read their comments and reply back. This is the case, for example, if they stumble over a mistake in an article's text, which could be a simple typo or wrong information. Other comments are questions to the journalists, for example, asking for background information on a topic. A study found that the majority of readers want journalists to engage more in the online comment sections [34]. For example, 61 percent of the readers would like journalists to post comments to clarify factual questions.

The majority of comments address the broader audience of all readers. Still, journalists can foster respectful discussions by joining the discussion as moderators. For highly controversial topics or when a discussion drifts to a disrespectful tone, their intervention could ensure compliance with the platform's rules. However, the sheer number of comments makes it infeasible for journalists to read each and every comment. As a consequence, they find comments that are interesting for them only once in a while and are not aware of all those that would require a reaction in addition. Interesting ideas get lost, discussions get out of focus, and journalists and readers get disappointed.

A first requirement to increase journalist engagement is to make them aware of the comments that require a reaction. We define these comments that are worth reading for journalists as relevant comments. They could be worth reading for different reasons: praise or criticism of the article or journalist, which does not necessarily require a reaction, or direct questions to the journalists, which should be answered. In this paper, we present a dataset of 38,000 English-language reader comments and 19,000 journalist replies from The Guardian to investigate the interactions of journalists with their readership. Half of the reader comments did receive one of the replies while the other half did not receive a reply by a journalist. The dataset comprises comment texts and metadata and is enriched with machine labeling. A sample of the dataset was also manually labeled. The labels address the reasons why a reader comment received a reply.

Further, we propose a novel task of recommending relevant reader comments to journalists. This task can either be interpreted as a classification or a ranking task. For the former, given a set of

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

reader comments, all comments that require a reply by a journalist need to be identified. For the latter, a given set of reader comments needs to be sorted by the necessity or likeliness of a reply by a journalist. As a baseline, we present a neural network model that is trained on a pairwise-ranking task. In an application scenario, a recommender system could show the most relevant comments to journalists together with an explanation of what makes them relevant. For example, journalists could get to see a personalized view that differs from the public view. Instead of displaying the comments in chronological order, the comments would be ranked based on each comment's relevance to the journalists. Thus, they could find the relevant ones at the very top and could reply to them quickly if necessary. If time permits, more comments could be explored in the order of their estimated relevance.

Article Outline. In the remainder of this paper, we start with a summary of related work and put our new dataset into context. After that, we describe the data and report baseline results for a comment ranking task. The dataset and the code for the experiment are published online for research purposes.[1]

2 RELATED WORK

Several comment datasets are publicly available for research purposes. The majority of them come from shared tasks on toxic comment classification, e.g. with focuses on hate speech against immigrants and women [5], offensive language [35, 40], and aggression [7]. They cover a diverse set of languages besides English: Spanish [15], Italian [9], Hindi [7, 21], Bangla [7], German [35, 39], Arabic, Danish, Greek, and Turkish [41]. Further, there is a large comment dataset from a Kaggle data science challenge[2] and several datasets of tweets that contain hate speech [12, 37]. Other comment datasets focus on what makes online discussions fruitful, and, for example, label "desirable content", such as "personal stories" or "rational argumentation" [32]. In the same line, the Yahoo News Annotated Comment Corpus is labeled with the goal to identify "engaging, respectful, and/or informative conversations" [25] and the SFU Opinion and Comments Corpus contains labels with regard to the "constructiveness" and the "toxicity" of comments [20]. Because of the immense labeling effort and the enormous amount of data needed for training deep neural networks, augmentation of comment datasets is a further line of work [28, 31].

The main body of related work on comment datasets investigates how to improve online discussions. One way to do this is by supporting content moderators and community managers through (semi-)automation of manual tasks. For example, one task for moderators is to identify high-quality comments — so-called "editor picks" and increase their visibility. The latter can be achieved by highlighting them or displaying them at the very top of the discussion section on a web page. Napoles et al. [25] define what makes a comment engaging, respectful, and informative. Based on an annotated dataset of 2,300 comment threads, they train a logistic regression classifier to detect such high-quality threads automatically [24].

Similarly, Kolhatkar and Taboada [19] classify constructive comments. They use editor picks from New York Times comments as positive examples, and comments from non-constructive threads from the Yahoo News Annotated Comments Corpus as negative examples to train their bidirectional, long short-term memory (LSTM) model. However, using training data from two different platforms is a potential source of error. In our approach, we focus on one platform at a time and are carefully sampling a non-biased subset from this data. Further, while Kolhatkar and Taboada aim to engage more readers, we aim to engage more authors of news articles to join a conversation.

Jaech et al. [18] evaluate the ranking of comments based on their "karma", which they define as the number of upvotes minus the number of downvotes a comment received. While they work on the platform Reddit, which is not exactly a news platform, the task is still similar to ours. Informativeness, relevance, and user reputation are identified to be essential features for finding high-karma comments. However, the importance of these features depends on the community. A combination of received upvotes and downvotes has also been used as an indicator for a comment's success and popularity in the community [22]. Other work supports moderators at finding high-quality comments with interactive applications [26] or with a crowd-sourced model for moderation [23]. There, the workload of content moderation is distributed among many users.

Diakopoulos and Naaman [13] analyze reasons why people read or write online news comments. Expressing opinions and emotions is the strongest motivation for writing comments, followed by the desire to provide information to others, such as answering questions, sharing experiences or correcting errors. On the reader's side, the strongest motive is learning about other readers' views. A similar analysis focuses on readers' motivation for commenting on news articles and blogs [4]. The authors aim to recommend "comment-worthy" articles to readers based on the article and its received comments. Recent research also focuses on how user engagement is affected by a threaded or a linear presentation of the comments in a user interface [3], and how the engagement varies by topic and news platform [1].

Forecasting the popularity of news articles helps moderation teams to schedule their workload [2, 33]. Popularity is measured in terms of the expected number of received comments because content moderators at the considered platforms need to check each and every comment for toxic content. Toxic comments are defined as comments that make other users leave a discussion. For example, insults, threats, and hate speech fall into this category. Semi-automated comment moderation is one recent outcome in this field of applied research [27, 29]. Modeling the dynamics of conflict in online discussions can also support moderators in deciding when to intervene [14].

Besides identifying toxic or high-quality comments and deleting or highlighting them by means of moderation, a third option to foster respectful discussions is to engage article authors in the discussion. This task drove the research in this paper and pointed out the lack of suitable training or test datasets. Closests to the kind of data needed for this task is the dataset provided by Schabus, Skowron, and Trapp [32]. It consists of almost 12,000 German-language news comments that have been labeled with regard to nine categories. One of their labeled categories is "feedback", which includes questions and suggestions addressing the article authors.

[1] https://hpi.de/naumann/projects/repeatability/text-mining.html
[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Table 1: Dataset Statistics

Comments	56 631
Articles	4 563
Readers	18 084
Journalists	432
Reader comments with/without journalist reply	18 877 / 18 877
Min/Median/Max comment length (chars)	1 / 239 / 4 985
Journalist replies	18 877
Min/Median/Max reply length (chars)	1 / 151 / 4 542

A subset of comments in this category might require a reply from the journalist or the editor.

There is also research on a feature-based approach for identifying comments that address a journalist or the news platform provider in general [16]. This research considers, for example, the section of the article as a feature but also the timestamp. In contrast, we focus on the comment text only and do not aim to predict journalist replies but to recommend what comments need a reply. This difference is significant because we do not want to predict the journalist replies as they were in the past, but we want to increase journalist engagement. Therefore, a good predictor that takes, e.g., the timestamp of a comment into account as a feature does not help in our scenario. We argue that our task is related to but different from prediction. While the task of comment ranking in general is not new, it is still an open research question how exactly to perform this ranking and what criteria to use [6, 11, 17, 36, 38]. To the best of our knowledge, no research has been conducted on the task of recommending reader comments to journalists so far. Further, there is no publicly available dataset that could readily be used for such research.

3 DATASET

The website of The Guardian reaches 22 million monthly readers making it the most popular online newspaper in the UK.[3] The dataset that we present is a subset of all reader comments that have been posted on this website. It investigates the interactions of the journalists with their readership. Half of the reader comments in our dataset received a reply by a journalist, while the other half did not. Due to the provided class labels and the balanced class distribution, the comments can be easily used for supervised machine learning. In addition, we include the journalists' replies for reference. The following section gives more details about the dataset and explains how it was collected. Statistics of the dataset are given in Table 1.

3.1 Collecting Reader Comments and Journalist Replies

We collected 51 million reader comments from The Guardian posted between November 2011 and December 2018. Before that time, there was simply no option to post a comment in the form of a reply to another reader's comment on this platform.[4] We selected all 18,877 reader comments that received a reply by the journalist (positive samples) and selected 18,877 reader comments that did not receive a reply by the journalist (negative samples). The negative samples are randomly sampled reader comments that did not receive a reply but were posted in a short time window before and after the journalist reply itself. This time window starts one hour before the journalist reply and ends one hour after it. Thereby, we can assume that the journalist was active on the platform and could have chosen to react to the negative sample. The journalists can be identified because the profile names of their official accounts match with the article author names. However, it could be that in rare cases the journalists use their private accounts to post comments under a pseudonym. We neglect these rare cases and focus on the official replies by journalists.

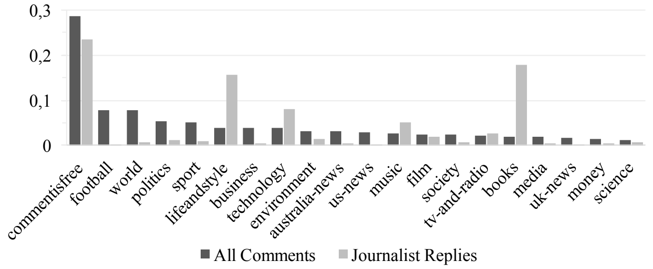

Figure 1: For the top 20 article sections of The Guardian, dark bars show the relative number of comments by readers and light bars show the relative number of journalist replies. For example, journalists engage especially in the books section.

For each considered news article, we sample the same number of positive and negative samples. Thereby, we prevent an article bias and ensure the same number of positive and negative samples per topic. Figure 1 shows the relative number of comments in the top 20 sections. The majority of all reader comments and also of all journalist replies are posted in the "comment is free" section. This section is where The Guardian main commentators and selected contributors from outside publish opinion articles. Interestingly, journalist replies peek for the sections "life and style", "technology", "music", and "books". Journalists seem to be more active in these sections: For example, while articles in the books section receive only two percent of all comments, they receive 18 percent of all journalist replies.

Figure 2 shows our database schema. Each comment in our dataset is represented with a comment id and its text. Further, the user id, timestamp, number of upvotes, and corresponding article URL are given. The latter includes the article's section, such as politics, sports or lifestyle. Those comments that are a reply to another comment reference their parent with its comment id. For each article, we know the user name of the journalist's official account and can thereby infer the user id. After matching articles with the user id of their authors, the user names are discarded. Last but not least, the dataset also contains the 18,877 replies by journalists.

3.2 Enriching the Data with Machine Labeling

We enrich the dataset with additional labels generated by a pretrained machine learning model. This model classifies whether a

[3] https://www.theguardian.com/

[4] https://www.theguardian.com/help/insideguardian/2011/nov/03/responses-in-comments

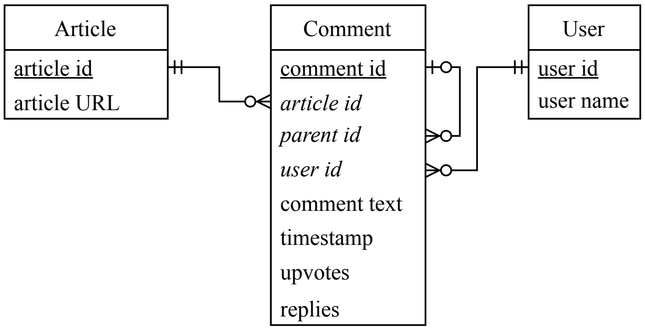

Figure 2: The dataset consists of commments from The Guardian, including their metadata, such as references to news articles or other comments if they are posted as a reply.

Table 2: Mean and Variance of Machine-Labeled Scores

	Positive Samples		Negative Samples	
Label	Mean	Var	Mean	Var
persuasive	38.3	2.4	35.8	2.4
audience	78.7	0.5	79.5	0.5
agreement	19.4	0.1	20.0	0.1
informative	37.0	1.6	35.1	1.5
mean	35.4	0.2	35.7	0.2
controversial	61.1	0.7	60.0	0.8
disagreement	60.5	0.3	60.5	0.3
off topic	57.7	0.6	58.7	0.6
neutral sentiment	46.0	0.1	46.1	0.1
positive sentiment	12.9	0.0	13.2	0.1
negative sentiment	69.3	0.1	69.3	0.1
mixed sentiment	32.9	0.9	31.5	0.9

comment is, e.g., informative or controversial. As a result it allows investigating whether informative reader comments are more likely to receive a reply from the journalist compared to controversial ones. We trained this model using the Yahoo News Annotated Comments Corpus [25]. It contains more than 9,000 comments that have been labeled with respect to twelve classes: persuasive, audience, agreement with commenter, informative, mean, controversial, disagreement with commenter, off topic with article, sentiment neutral, sentiment positive, sentiment negative, and sentiment mixed. For a detailed description of the classes, see the paper by Napoles et al. [25]. Following their approach, we train a logistic regression classifier on this data to automatically label comments. Thereby, the dataset is enriched with a machine labeling approach.

Table 2 lists the mean label scores for the set of comments that received a reply by the journalist (positive samples) and the set of comments that did not receive a reply by the journalist (negative samples). The scores differ only slightly for the majority of labels. The label with the most significant difference is persuasiveness: the average score is 38.3 percent for the positive samples, while it is 35.8 percent for the negative samples. However, the variance is also the highest among all labels.

Analyzing the comments and the machine-labeled scores by hand, we find that journalists not only reply to comments where it is obvious that a reply is required but to random-looking comments. For example, a comment that only consists of the emoticon ":^)" is machine-labeled as 2.4 percent persuasive and 4.5 percent informative, but still received a reply by the journalist. Similarly, the comment "Sounds good to me…!" is machine-labeled 12.3 percent persuasive and 15.4 percent informative, but still received a reply by the journalist. An inherent limitation of the machine labeling approach is that it was trained on a different dataset and the classification is, therefore, more error-prone. For example, even if a comment is not informative, the model sometimes assigns a high probability to this label.

3.3 Manual Labeling Procedure

In addition to the machine labeling approach, we manually labeled the data with regard to the reasons why journalists replied to reader comments. The labels are based on a hierarchical taxonomy for user engagement in online news discussions [30]. Originally, this taxonomy was used for labeling the reasons why some comments receive an above-average number of replies and upvotes.
There are the following four main classes and fourteen subclasses:

Question Asking with the serious intent to receive a reply.

Explanation Expects an answer to the questions *Why? How can I...?* in the form of a (long) explanation.
Opinion Expects a long, subjective answer to the question *What do you think about...? What would you suggest?*
Fact Expects a short, objective answer to *When?* or *Where?*

Information Contributing new information to the discussion.

Correction Pointing out a (perceived) mistake in the article (spelling mistake/factual error) and/or suggesting a correction.
Personal Story Telling a background story, e.g., about the reader, another person or somebody's work place.
Fact Adding objective, factual information, such as a statistic.

Opinion Expressing a point of view and/or convincing others.

Consent with Comment/Article Topic Positive sentiment towards/agreement with another comment/the article topic.
Dissent with Comment/Article Topic Negative sentiment towards/disagreement with another comment/the article topic.
Suggestion Telling what needs to be done/changed.
Speculation Unverified statements about either about the **future** (predicting what will happen) or giving **reasons** as an assumed explanation, e.g., of an event or the actions of somebody.

Joke/Humor Making others laugh or making fun of somebody.

A subset of one thousand reader comments that received a reply by a journalist were labeled according to this taxonomy in a joint effort by three annotators in parallel. During the annotation process, the annotators had access to the reader comment, the journalist reply, and the title of the news article. On the coarse-grained level (with only the four classes question, information, joke/humor, and opinion), at least two out of three annotators agreed on the labels for 90 percent of the comments. The inter-annotator agreement in terms of Fleiss' Kappa is 0.42. On the fine-grained level, at least

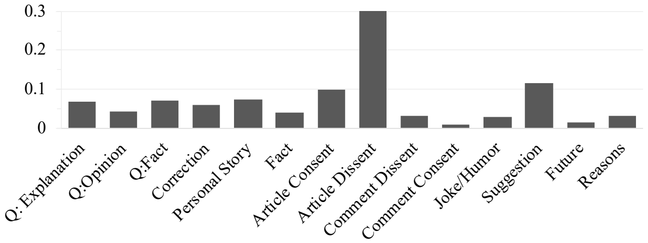

Figure 3: Distribution of the manually created class labels for the reader comments. The most frequent comments that receive a reply by the journalist are comments with negative sentiment towards the article topic. Corrections are at seventh rank. For a detailed description of the classes, see the related work [30].

two out of three annotators agreed on the labels for 70 percent of the comments. The inter-annotator agreement in terms of Fleiss' kappa is 0.37.

One collective label was derived from the individual labels by the annotators. To this end, in cases where two annotators agreed, their label decision overruled the third annotator. In cases where all annotators disagreed, we excluded the sample from the published dataset. Figure 3 shows the resulting label distribution. The majority of the comments that received a journalist reply were labeled as dissent with/negative sentiment towards the article text/topic. The labels serve as a starting point for investigating *why* journalists reply to particular reader comments.

3.4 Accessing the Dataset

The dataset can be downloaded via the official Web API of The Guardian. To this end, we provide predefined lists of 38,000 reader comments and 19,000 journalist replies identified by their comment IDs, and a small script that accesses the API. This script also joins the retrieved comment texts and metadata with our labels. An API key is required for using the API, which can be applied for via an online form by providing name and email address and accepting the terms and conditions of the Guardian Open Platform. On the one hand, the approach allows other researchers to re-create the dataset, repeat experiments, and continue research in this field. On the other hand, platform users still have the option to delete (or edit) their own comments so that they are not shown on the Website and cannot be retrieved via the API anymore.

3.5 Training Domin-Specific Word Embeddings

The dataset contains 38,000 labeled training instances (reply received/not received), which is enough to train a deep-learning-based neural architecture. As input, we use a word embedding representation of all comments. However, 38,000 short comments are not enough to train word embeddings on this dataset. Therefore, we pre-train 300-dimensional word embeddings on the full dataset of all 60 million comments (between 2006 and 2019) from The Guardian with the fastText method [8]. This corpus comprises 4.4 billion tokens, and its size is on the same level as Wikipedia corpora used to train word embeddings, which contain 2 to 4 billion tokens. The full text is transformed to lowercase, and user mentions and URLs are replaced with special tokens. The embeddings are trained for five epochs using the skip-gram method and subwords of three to six characters. After pre-training, the weights of the word embedding layer are kept fixed during the training of the full neural model. FastText is superior to other embedding methods, such as Word2Vec or GloVe, because of the benefits of subword embeddings [8]. They prevent out-of-vocabulary issues, which otherwise frequently occur with online comments. Typos and neologisms are not uncommon in this scenario. Figure 4 visualizes a two-dimensional projection of the embedding space with a focus on the nearest neighbors of the word *troll*. The trained word embeddings are published alongside the dataset.

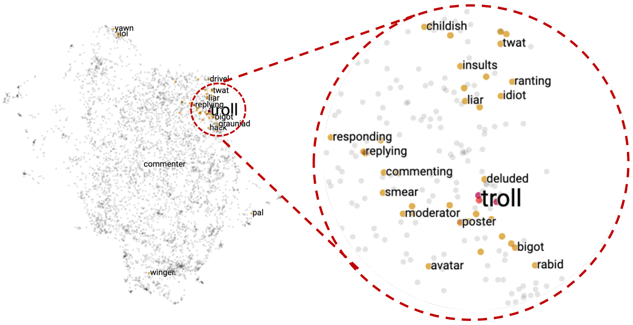

Figure 4: This UMAP projection of the domain-specific word embeddings trained on 60 million comments from The Guardian highlights the nearest neighbors of the word *troll*. In the high-dimensional space, *troll* is embedded close to *trolling, trolls, commenter, poster,* and several swear words.

4 BASELINE APPROACH AND EXPERIMENTS

The labels provided with the dataset allow investigating many aspects of the interactions between journalists and their readership, such as the dynamics of the sentiment of their comments, journalists explanations and apologies for mistakes in their articles, or correlations of the text length of a reply and the number of received upvotes by readers. In this paper, however, we focus on the task of identifying comments that require a reply by the journalist and ranking the comments accordingly. There is no clear borderline between comments that do or do not require a reply. While the journalist, in the end, has to make a binary decision (to reply or not to reply), the binary training data with positive and negative samples is only a small excerpt of reality. There is no single correct solution, and different journalists react for different reasons. Therefore, we consider a ranking task of comments instead of a binary classification task. We rank comments by the likelihood of receiving a reply so that journalists can get the most relevant comments displayed at the top of the discussion sections. We suggest generating a ranking of all comments step-by-step with a pairwise ranking approach. Given two comments, the task is to decide which one is more relevant to journalists, i.e., which comment is more likely to receive a reply by them. The pairwise decisions can be aggregated with a variety of methods to obtain a global ranking.

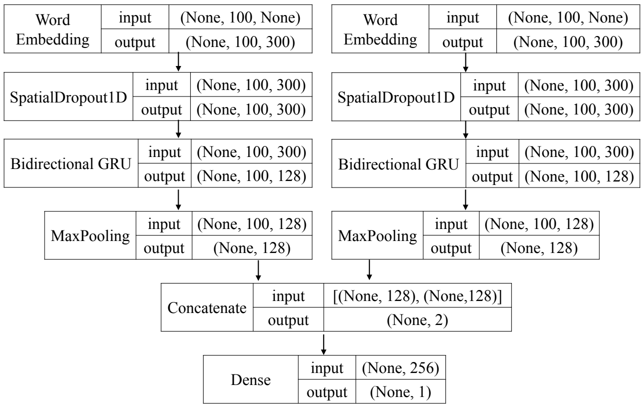

Figure 5: A Siamese neural network architecture allows a pairwise ranking of comments.

To solve the described task, we train a deep neural network on pairs of positive and negative samples. The input of the neural network consists of two comments, where one received a reply by the journalist, and the other did not. Figure 5 visualizes the network's architecture. The network exhibits a Siamese structure where the two inputs pass through an encoder of two word embedding layers and two bidirectional gated recurrent unit (GRU) [10] layers that run in parallel and share weights. The two bidirectional GRU layers each encode the sequence of word embeddings of a comment, and each of them is followed by a max-pooling layer. The output of the two max-pooling layers is concatenated. The final output of the network is calculated by a dense layer with a single sigmoid activation function and describes which of the two input comments should more likely receive a reply. For training the network, we use binary cross-entropy as the loss function and an Adam optimizer. Not considering the word embedding layer, the neural network has a rather simple structure with a small number of trainable parameters. This limited capacity is tailored to the comparably small size of the training dataset.

For the evaluation of the model, the dataset is split into 80 percent training, 10 percent validation, and 10 percent test data. The comment IDs for these splits are published to foster repeatability and set a standard for comparisons in future work. To tune the number of training epochs, we use early stopping and monitor the loss on the validation set. Because of the balanced class distribution, the evaluation uses classification accuracy. The model achieves an accuracy of 64.0 percent, which means that the two input comments are ranked in the correct order in about two-thirds of the cases — leaving room for improvement.

5 DISCUSSION

Our work provides an insight into how the media and publishing industry can benefit from the ranking and classification of reader comments. However, the studied machine learning approaches demand labeled training data, which is costly to obtain. Identifying and gathering this critical component appears to be a major challenge.

Recommending journalists when to reply to a reader comment is a hard task, not least because it finally comes down to a journalist's personal decision. Some journalists are more active on the platform than others. There are several reasons for this variety. First, journalists work in different sections, such as politics, sports, books, and therefore, the discussion topics differ. Some topics are suited for journalists to post their personal opinion, for example, when it comes to book recommendations. Other topics demand strict neutrality, such as football matches, where readers support opposing teams. Further, journalists might be unavailable at the time of publication of their article and therefore, cannot reply to reader comments in time. Because of the short attention span and fast-paced media business, a journalist reply that is posted a few hours later would go mostly unnoticed by users. The presented baseline approach neglects that journalists have different notions of what makes comments worth reading and worth replying to.

5.1 Outliers and Noise

Given the subjective nature of the task of replying to a reader comment, there are some outliers in our dataset that would be hard (if not impossible) to predict. Sometimes even very short comments gained the journalists' attention and resulted in a reply. For example, the short comment "Yay!" received a reply by a journalist:

> **reader A:** I'm sure Gunny would plug for Tom Tomorrow, and i'd plug for Jen Sorensen too.
>
> ⋮
>
> **reader A:** Yay!
>
> > **journalist:** Evidently, I should get acquainted with Jen Sorensen's work also [. . .]

The reason for this lies in the other comments and their context, specifically in a comment posted by the same user earlier. Actually, the journalist replied to both comments by the reader but the dataset indicates it only as a reply to the shorter comment.[5] The comments are part of a discussion about layout and feature changes of the comment section, which explains why the journalist more actively joins the discussion: The discussion topic is THE GUARDIAN itself. The machine-labeled probabilities of persuasiveness and informativeness are low and fail to identify the reader comment as relevant to the journalist.

Another example for a hard to predict reply stems from an article entitled "How to eat: beef stew".[6]

> **reader B:** No mention of Yorkshire Pudding as an accompaniment? An absolute must in our house.
>
> > **reader C:** Instead of dumplings or instead of bread?
> >
> > > **journalist:** There is, of course, nothing that Yorkshiremen wouldn't eat out of a pudding. Trifle, cereal, you name it.

The reply by reader C is machine-labeled with low persuasiveness (0.15) and informativeness scores (0.18), but, surprisingly, the journalist replied to it. This reply is unpredictable for our model and it is a debatable point whether this sample helps the training process or rather is noise that should be removed in a preprocessing step.

[5] https://gu.com/help/insideguardian/2012/apr/23/makeover-comment-is-free-america

[6] https://gu.com/lifeandstyle/wordofmouth/2014/feb/27/beef-stew-bread-dumplings-bowl-no-potatoes

5.2 Industry Impact

Through our regular meetings with journalists, it became clear that traditional one-way journalism is on the decline. Participatory news platforms are more and more a space for discussion and exchange. Today, providing news to the public is not enough. Readers expect a news platform not only to present facts and individual opinions but to enable dialogs, discussions, and sharing of their own viewpoint. In addition, readers writing comments want to be noticed and expect journalists to answer questions or engage in discussions. With our dataset and approach, we aim to analyze and facilitate the interaction of readers and journalists by dealing with the massive amount of daily comments a news article receives. Detecting which comment necessitates a reaction by the journalist or asking readers to manually indicate if and why that is the case will most likely increase the interaction on the platform by showing relevant comments to the journalist. This personal contact between journalists and commenters strengthens the sense of community and thus enhances the user experience. We hope that this is one step towards more respectful conversations and engaging discussions in online forums. By fostering this sense of community, we hope to mitigate hate speech, insults, and disrespectful utterances. And eventually, through the effects of group dynamics and self-regulation, we hope to avoid the automatic deletion of offending comments and perceived censorship.

6 CONCLUSIONS AND FUTURE WORK

We presented a new data set consisting of 38,000 reader comments and 19,000 journalist replies posted to the website of The Guardian. Thereby, we enable research on the novel task of recommending reply-worthy reader comments on news platforms to journalists. It is unfeasible for them to keep track of all received comments and decide when to reply. However, readers sometimes expect a reply, for example, if they ask questions addressing the journalists or point out mistakes in the news articles. This dilemma motivates to investigate the engagement of journalists with their readership. To this end, the dataset is labeled with regard to the causes that triggered replies by the journalists. A deep neural network serves as a baseline approach for a pairwise ranking task. Journalists could apply this model to re-rank the comments so that the most relevant comments are shown at the top of the comment sections (in contrast to the standard chronological ranking).

There are several promising paths for future work. The most critical challenge to tackle is interpretability. Right now, users cannot comprehend reasons for a particular ranking, and it remains unclear how different rankings compare to each other with regard to user relevance. While our baseline ranking reliably removes obviously irrelevant comments, the correct ranking of two relevant comments is not well-defined. Therefore, an evaluation of comment rankings with a user study is needed. With the continuing development of comment ranking tools, we envision that comment sections on online news platforms become more relevant to readers and journalists. Thereby, we hope to increase engagement and foster meaningful discussions that complement news articles.

ACKNOWLEDGMENTS

First of all, we would like to thank The Guardian for providing access to the comment data via their Guardian Open Platform. This project was partly funded by the Robert Bosch Foundation in the context of its master class "Science Journalism" 2018. We would like to thank the collaborating data scientist Andreas Loos from ZEIT ONLINE and journalist Wolfgang Richter from Journalistenbüro Schnittstelle for the close cooperation. Further, we would like to thank our students Nicolas Alder, Lasse Kohlmeyer, and Niklas Köhnecke for their support throughout the project.

REFERENCES

[1] Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. 2019. View, Like, Comment, Post: Analyzing User Engagement by Topic at 4 Levels across 5 Social Media Platforms for 53 News Organizations. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, Vol. 13. AAAI, 47–57. Issue 1.

[2] Carl Ambroselli, Julian Risch, Ralf Krestel, and Andreas Loos. 2018. Prediction for the Newsroom: Which Articles Will Get the Most Comments?. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. ACL, 193–199.

[3] Pablo Aragón, Vicenç Gómez, and Andreaks Kaltenbrunner. 2017. To thread or not to thread: The impact of conversation threading on online discussion. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*. AAAI, 12–21.

[4] Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. 2015. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *Proceedings of the Conference on Recommender Systems (RecSys)*. ACM, 195–202.

[5] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval@NAACL)*. ACL, 54–63.

[6] 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)*. International World Wide Web Conferences Steering Committee, 1371–1380.

[7] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@LREC)*. 158–168.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)* 5, 1 (2017), 135–146.

[9] Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Vol. 2263. 1–9.

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 1724–1734.

[11] Onkar Dalal, Srinivasan H Sengemedu, and Subhajit Sanyal. 2012. Multi-objective ranking of comments on web. In *Proceedings of the International Conference on World Wide Web (WWW)*. ACM, 419–428.

[12] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*. AAAI, 512–515.

[13] Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 133–142.

[14] Subhabrata Dutta, Dipankar Das, Gunkirat Kaur, Shreyans Mongia, Arpan Mukherjee, and Tanmoy Chakraborty. 2019. Into the Battlefield: Quantifying and Modeling Intra-community Conflicts in Online Discussion. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*. ACM, 1271–1280.

[15] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval@SEPLN)*. CEUR, 214–228.

[16] Marlo Häring, Wiebke Loosen, and Walid Maalej. 2018. Who is Addressed in This Comment?: Automatically Classifying Meta-Comments in News Comments.

Proceedings on Human-Computer Interaction (HCI) 2, CSCW, Article 67 (Nov. 2018), 20 pages.

[17] Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. 2009. Ranking comments on the social web. In *Proceedings of the International Conference on Computational Science and Engineering (CSE)*, Vol. 4. IEEE, 90–97.

[18] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions?. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2026–2031.

[19] Varada Kolhatkar and Maite Taboada. 2017. Using New York Times Picks to Identify Constructive Comments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 100–105.

[20] Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2019. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics* 4, 2 (2019), 155–190.

[21] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*. ACL, 1–11.

[22] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*. AAAI, 311–320.

[23] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 543–550.

[24] Courtney Napoles, Aasish Pappu, and Joel R Tetreault. 2017. Automatically Identifying Good Conversations Online (Yes, They Do Exist!). In *Proceedings of the International Conference on Web and Social Media (ICWSM)*. AAAI, 628–631.

[25] Courtney Napoles, Joel R. Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the Linguistic Annotation Workshop (LAW@EACL)*. 13–23.

[26] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 1114–1125.

[27] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep Learning for User Comment Moderation. In *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*. ACL, 25–35.

[28] Julian Risch and Ralf Krestel. 2018. Aggression Identification Using Deep Learning and Data Augmentation. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*. ACL, 150–158.

[29] Julian Risch and Ralf Krestel. 2018. Delete or not Delete? Semi-Automatic Comment Moderation for the Newsroom. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*. ACL, 166–176.

[30] Julian Risch and Ralf Krestel. 2020. Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, 579–589.

[31] Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*. ACM, 991–1000.

[32] Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 1241–1244.

[33] Erez Shmueli, Amit Kagian, Yehuda Koren, and Ronny Lempel. 2012. Care to Comment?: Recommendations for Commenting on News Stories. In *Proceedings of the Conference on World Wide Web (WWW)*. ACM, 429–438.

[34] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. 2016. News commenters and news comment readers. *Engaging News Project* (2016), 1–21.

[35] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*. German Society for Computational Linguistics & Language Technology, 352–363.

[36] Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2016. Learning to rank non-factoid answers: Comment selection in web forums. In *Proceedings of the International on Conference on Information and Knowledge Management (CIKM)*. ACM, 2049–2052.

[37] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the Workshop on NLP and Computational Social Science (NLP+CSS@EMNLP)*. ACL, 138–142.

[38] Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 2. ACL, 195–200.

[39] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*. Austrian Academy of Sciences, 1–10.

[40] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the International Workshop on Semantic Evaluation (SemEval@NAACL)*. ACL, 75–86.

[41] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval). In *Proceedings of the International Workshop on Semantic Evaluation (SemEval@COLING)*. ACL, 17.