Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions

Julian Risch, Ralf Krestel
Hasso Plattner Institute, University of Potsdam, Germany
firstname.lastname@hpi.de

Abstract

Comment sections below online news articles enjoy growing popularity among readers. However, the overwhelming number of comments makes it infeasible for the average news consumer to read all of them and hinders engaging discussions. Most platforms display comments in chronological order, which neglects that some of them are more relevant to users and are better conversation starters.

In this paper, we systematically analyze user engagement in the form of the upvotes and replies that a comment receives. Based on comment texts, we train a model to distinguish comments that have either a high or low chance of receiving many upvotes and replies. Our evaluation on user comments from *TheGuardian.com* compares recurrent and convolutional neural network models, and a traditional feature-based classifier. Further, we investigate what makes some comments more engaging than others. To this end, we identify engagement triggers and arrange them in a taxonomy. Explanation methods for neural networks reveal which input words have the strongest influence on our model's predictions. In addition, we evaluate on a dataset of product reviews, which exhibit similar properties as user comments, such as featuring upvotes for helpfulness.

User Comments in Online News Discussions

Thirty years ago, newspapers received hand-written letters to the editor and selected maybe a handful for publication. This was called reader engagement and was the only way for readers to interact with other readers and/or the newspaper via public discussion. With the rise of the World Wide Web, the establishment of online news platforms, and the appearance of online discussion sections, the situation has changed drastically. Nowadays, irrespective of who the readers are and what they think, they can exercise their right to freedom of speech and freely share their opinion.

On the flip side, the ever-increasing number of comments not only distracts readers, but also hinders engagement. No news consumer is able to read through all the comments. Overwhelmed by hundreds to thousands of comments, new users give up on joining the discussion. A current approach for coping with this information overload is to highlight comments that are especially interesting in the eyes of the editors. This manual effort is costly and comes on top of the task of moderating hate speech and other banned content.

Major news platforms allow users to upvote comments, but for several reasons these platforms do not use votes as a ranking criterion for comments. First, there is the cold start problem: Whenever a new comment is posted, it has not yet received any upvotes. An accordingly low rank affects the comment's exposure to users and reduces its chance of ever receiving any upvotes. Moreover, such a ranking algorithm can easily be gamed. Malicious users can register multiple accounts or collaborate to break the ranking system and upvote comments of their favored opinion.

Today's platforms refrain from using an upvote-based ranking algorithm and simply sort comments chronologically. They give no incentive for the described manipulations. Thus, casting an upvote conveys a sense of relevance to the respective user — but nothing more. A comment receives many upvotes if it motivates many users to engage by voting for it. We make use of this information to build a dataset of comments that are either most or least engaging. Note that some platforms also allow users to downvote comments, which we do not analyze in our work.

Voting on a comment is a rather basic way to interact. In contrast, replying to another user's comment actually starts a conversation. Users reply to comments for different reasons. For example, they want to correct another user's error, give their personal view, or express consent or dissent. While the number of upvotes reveals a comment's popularity, the same does not hold for the number of replies. Besides upvotes, we also consider replies as a form of user engagement and give insights into their interplay.

In this paper, we classify engaging comments without costly manual annotation effort by editors. Therefore, we leverage user reactions that are inherent to online discussions: comment upvotes and replies. The number of these reactions distinguishes the most and the least engaging comments, which we also refer to as top comments and flop comments. In our experiments we analyze a real-world dataset of user comments from the British online news platform *TheGuardian.com.* Figure 1 is a screenshot of the platform's comment section. For illustration purposes, we list two comments that generated a large amount of engagement in the form of many upvotes or replies:

1. "The brexiters are achieving their wish: they're turning

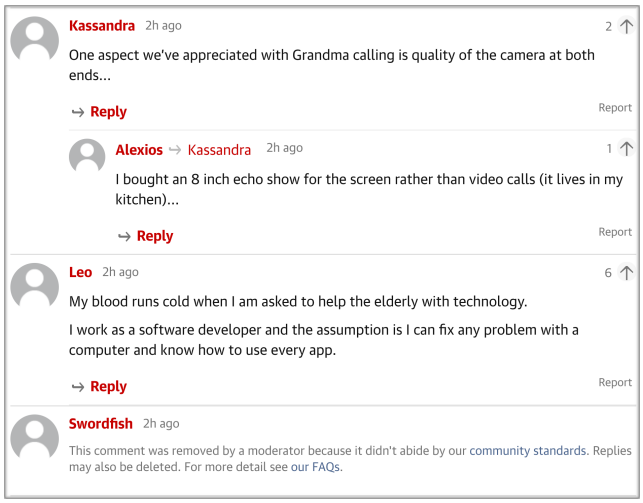


Figure 1: On *TheGuardian.com*, readers can post a comment, cast an upvote, and reply to another user's comment.

the UK into the kind of second rate country they can feel at home in." 2,615 upvotes, 3 replies

2. "Can somebody please explain to me why some people are so rabidly anti-gay marriage?" 82 upvotes, 20 replies

The first comment refers to an anticipated loss of 1,000 jobs in EU authorities located in the UK as a consequence of Brexit. The number of received upvotes is extraordinarily high, presumably because many anti-brexiters identify with the expressed opinion. The second comment was posted half a year before the UK parliament legalized same-sex marriage. It was a topic of controversial discussions at that time with replies containing different opinions and address the user's request for explanation.

Being able to automatically detect such engaging comments is important for many applications. The most obvious one would be a ranking criterion to display user comments from most engaging to least engaging. Another application field would be in the context of recommending comments to readers, either to reply or to simple read them. Currently this recommendation is done manually through editor picks. Finally, users writing many engaging comments could be rewarded by the news platform. This would further incentivize users to write high quality comments.

Contributions In summary, the contributions of this paper are: (1) defining user engagement in online discussions based on adjusted number of upvotes and replies; (2) designing a taxonomy that explains engagement triggers; (3) proposing a neural network model to distinguish most and least engaging comments based on their text; (4) evaluating classification accuracy of the proposed model on two datasets: user comments and product reviews. Our implementation, the evaluation datasets, and two models of domain-specific word embeddings are published online.[1]

[1]http://hpi.de/naumann/projects/repeatability/text-mining.html

Related Work

A growing body of research aims to foster respectful and fruitful discussions on the Web. Applications of this research manifest in real-world system implementations that support moderators and community managers, for example, by predicting how many comments a news article will receive (Ambroselli et al. 2018), identifying comments that require moderation (Schabus and Skowron 2018; Risch and Krestel 2018) or highlighting comments that are worth reading (Park et al. 2016). To this end, there are two primary directions of related work on comment classification: identifying either toxic or high-quality comments.

The term *toxic comments* comprises hate speech, insults, threats, profanity, and content that otherwise makes users leave a discussion. Platforms enforce a ban on toxic comments through manual moderation, but the ever-increasing number of comments renders this effort infeasible. Several studies work towards automation of this step and train deep neural networks on large datasets of annotated comments (Nobata et al. 2016; Wulczyn, Thain, and Dixon 2017; Badjatiya et al. 2017). The preparation and analysis of such datasets is a complex research task on its own (Schabus, Skowron, and Trapp 2017; Chen, Mckeever, and Delany 2017). A significant challenge is the inherent class imbalance of the data: typically, less than five percent of the comments are toxic (Xu et al. 2012; Risch and Krestel 2018). Recently, it has been proposed to not only classify single comments but also predict whether the tone of a sequence of comments is getting out of hand (Zhang et al. 2018).

Highlighting high-quality comments is the complementary task to deleting toxic comments. Related work defines the notion of quality in different ways, varying from *engaging, respectful, and informative* (Napoles, Pappu, and Tetreault 2017) to *interesting or thoughtful* (Diakopoulos 2015) and *constructive* (Kolhatkar and Taboada 2017). Similar to the task of toxic comment classification, state-of-the-art approaches use supervised machine-learning and require large amounts of training data, e.g., 2.3k annotated conversations (Napoles et al. 2017) or 30k annotated comments (Kolhatkar and Taboada 2017). We refrain from costly annotation efforts in our work and instead draw on information inherent to the data: upvotes and replies by users.

Kolhatkar and Taboada (2017) use editor picks from the *New York Times* as positive training samples to learn to identify constructive comments. These picks are a selection of comments judged as interesting or thoughtful by news editors. Negative training samples are taken from *Yahoo News* comments that were annotated as non-constructive in previous work (Napoles et al. 2017). Lampe and Resnick (2004) find that users generally agree on what comments are of high or low quality. However, users pay more attention to the earliest comments and top-level comments in a conversation than to responses. This finding motivates us to identify and remove this position bias in our dataset. Consequently, the number of upvotes and replies does not depend on the comment's position in the chronological ranking anymore. We refer to this position as the comment's rank in the following.

Online discussions are also mined to predict popularity of news stories (Rizos, Papadopoulos, and Kompatsiaris

2016), measure how controversial a comment is (Gómez, Kaltenbrunner, and López 2008), or rank comments by persuasiveness (Wei, Liu, and Li 2016). Hsu, Khabiri, and Caverlee (2009) make use of upvotes to rank comments, which is similar to parts of our approach. They measure a comment's visibility (exposure to users) by considering the popularity of the corresponding news article and the time between the publication of the article and the comment. Inspired by this idea, we use a comment's position in the chronological ranking to account for its visibility.

Related to our work, there is research on conversation modeling (Kumar, Mahdian, and McGlohon 2010; Wang, Ye, and Huberman 2012; Gómez et al. 2013; Backstrom et al. 2013; Aragón et al. 2017; Medvedev, Delvenne, and Lambiotte 2019) and on the dynamics of re-tweets (Zhang et al. 2016; Kobayashi and Lambiotte 2016). However, the motivation behind re-tweeting is to spread information in a social network, and in this regard it differs from replies in news discussions. The reasons users post comments on news articles are manifold. They range from expressing an opinion, asking questions, and correcting factual errors, to giving misinformation with the intent of seeing the community's reaction (Diakopoulos and Naaman 2011). We propose a taxonomy to characterize the different kinds of comments that trigger engagement by other users.

Berry and Taylor (2017) study the ranking of posts on public Facebook pages. They compare chronological ranking to ranking via social feedback and find that the latter has a positive effect on response quality. This insight motivates further research on ranking criteria for online comments aside from chronological ranking.

Most related work refrains from using upvotes as a feature, because of the many different factors that influence the number of upvotes a comment receives, such as its rank. At least, the interplay of a post's title, text, and publication time to predict user votes on Reddit and YouTube have been subject to research (Lakkaraju, McAuley, and Leskovec 2013; Siersdorfer et al. 2010). Chronological ranking in discussion threads is an essential difference in news comments compared to, for example, posts on Twitter or Facebook that can stand alone without a conversational context. In contrast to related work that predicts the popularity of a news article and the number of received user comments (Ambroselli et al. 2018), we predict the users' interactions with a comment. To this end, we neglect the news article text and focus on the comment text, upvotes, and replies.

Characterizing Users and Comments

In this section, we introduce and analyze a dataset of user comments from *TheGuardian.com*. There is an inherent position bias in the number of upvotes and replies and, after removing this bias, we find that top and flop comments differ in their average length and sentiment. Further, we visualize words that occur more often in either top or flop comments. Last but not least, we introduce a taxonomy to systematically categorize different types of engaging comments.

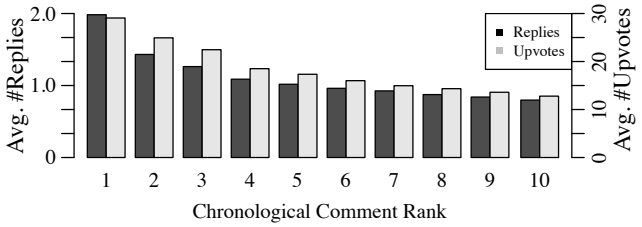


Figure 2: A comment's average number of received upvotes and replies correlates with its chronological rank in the discussion thread. This correlation is called position bias.

User Comment Dataset

The dataset comprises 61 million comments posted between 2006 and 2018. 1.2 million users contributed them in discussions of 600k news articles. For each comment, there is the comment text, user name, publication timestamp, corresponding news article, upvotes, and parent comment (if applicable). Half of the comments (53 percent) are replies to another user's comment and thus have a reference to this parent. Before November 2011, there was no option to post a reply in reference to another users comment on *TheGuardian.com* (Belam 2011). Therefore, we limit the dataset to the time after 2011 whenever we study the replies. We neglect that there was another change on the platform in 2012, when a single-level threaded design was introduced for the comment section (Hanman 2012). Budak et al. (2017) analyze the impact of this change on users and their discussions.

Upvotes cover the full timespan from 2006 to 2018 so that there is no need to limit the dataset when we study them. There are 260 million upvotes in total. While we have no knowledge of when, why and from whom a particular comment received upvotes, we know its final number of upvotes. We identify a position bias in the upvotes: the number of upvotes and replies that a comment receives depends on its rank. Figure 2 visualizes this dependency and reveals that earlier comments receive more upvotes and replies on average. In line with related work (Hsu, Khabiri, and Caverlee 2009), we attribute the advantage of earlier comments to their greater exposure to more readers. For this reason, the raw upvote and reply count is not enough to judge a comment's relevance to users in comparison with other comments. That is why we propose an approach to normalize the counts and thereby prevent the position bias from distorting the results.

In short, this approach transforms the absolute counts to relative numbers and afterwards groups all comments by their rank. For example, we compare a comment at rank 3 to all comments that appeared in other discussions at the same rank 3. Let us assume that the comment received 20 percent of all upvotes on the ten first comments in its corresponding article discussion. If comments at rank 3 receive on average less than 20 percent of the upvotes, we have identified a top comment, otherwise a flop comment. We describe the approach in more detail in the section *Distinguishing Top and Flop Comments*.

Based on the approach, we distinguish between two sets

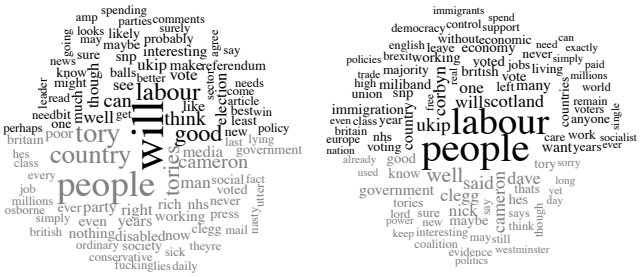
Table 1: The most and the least engaging comments differ in length and amount of neutral sentiment.

	Upvotes		Replies	
Average per Comment	Most	Least	Most	Least
Number of Words	75.54	43.68	76.82	38.52
Rate of Function Words	0.43	0.43	0.44	0.43
Rate of Personal Pronouns	0.13	0.12	0.12	0.13
Readability Index	9.82	9.08	9.50	9.14
Positive Sentiment	0.47	0.48	0.49	0.45
Neutral Sentiment	0.07	0.23	0.09	0.20
Negative Sentiment	0.46	0.30	0.42	0.34

of the most and least engaging (top and flop) comments and analyze their differences. As user engagement varies by news topic (Aldous, An, and Jansen 2019), we reduce the topical variety by limiting our analysis to comments on articles in the politics section. It is the section with the largest number of comments received. Table 1 compares the most and least engaging comments with regard to their average length, readability, and sentiment. Comments that generate less engagement are on average shorter and more often have a neutral sentiment. However, there is no difference in readability or the use of function words and personal pronouns. We use the automated readability index (ARI) to evaluate the readability. It is a standard metric that takes into account a text's characters per word and words per sentence.

Figure 3 compares the usage of the 100 most frequent words in comments that received the most or the least upvotes or replies. The word clouds display a word in the top half (black font) if it occurs more often in the most engaging comments and in the bottom half (gray font) if it occurs more often in the least engaging comments. The font size corresponds to the difference in the word's relative frequencies in both classes. For example, the relative frequency of the word *Labour* is 0.39 percent in comments that receive the most replies and 0.27 percent in comments that receive the least replies. The comparably large difference between these frequencies is illustrated by the word's large font size.

The most engaging comments mention the word *Labour* more often and the word *Tory* less often. The same relation holds for politicians of the respective parties, e.g., for Jeremy Corbyn (Labour) and David Cameron (Tory). A reason for this might be the political orientation of *TheGuardian.com* readers: according to a post-election survey, 73 percent voted for the Labour party and 8 percent for the Tory party in the 2017 UK general election (Curtis 2017). *TheGuardian.com* readers tend to upvote comments about their preferred party more often than comments about the opposite Tory party. This bias exemplifies why upvote counts cannot readily be used to distinguish high-quality from low-quality comments. Upvotes are cast with a subjective opinion in mind rather than with an objective and unbiased view of the comment text only. Another interesting example are comments that mention the word *people*. These comments receive few upvotes but many replies, probably because they make generalized claims about groups of



(a) Upvotes (b) Replies

Figure 3: Comparison word clouds show indicative words for classes of the most (black) and least (grey) engaging comments. For example, comments that mention *people* receive few upvotes but many replies.

people, which are controversial and serve as conversation starters. They are comparably unpopular on the platform but trigger many disapproving replies. Two examples are: "The people who voted for the war should be sent to prison as well." and "People are disillusioned with mainstream politics, and are starting to look elsewhere.".

Taxonomy of Engaging Comments

Different taxonomies have been proposed for hateful comments (Salminen et al. 2018; Waseem et al. 2017) but not for engaging comments. To foster a better understanding of engagement triggers, we propose a taxonomy for engaging comments, which is shown in Figure 4. We follow an open coding approach, also used by Salminen et al. (2018), and code 1500 engaging comments. With this approach, we organize classes in a conceptual hierarchy. Figure 5 exemplifies each class with a sample comment. For example, the class *Question* groups the subclasses *Explanation*, *Opinion*, and *Fact* together because all of them generate engagement by requesting answers in the form of comment replies. Comments in all three subclasses typically contain an exclamation mark. Note that the example comments for other classes, such as *Joke/Humor* and *Speculation* also contain questions. However, these questions are more of a rhetorical nature, and the corresponding comments trigger engagement for other reasons. The taxonomy also distinguishes between comments that trigger only upvotes, replies, or both. For example, while comments with jokes rarely receive replies, they frequently receive upvotes. It is the opposite if a comment asks for other users' opinions. However, if a comment dissents from a news article, other users express their approval or disapproval with both upvotes and comments. Our taxonomy is constructed in particular for comments on *TheGuardian.com* and is by no means universal. Other platforms might exhibit other classes of engaging comments, for example, if they allow users also to downvote comments. We revisit our taxonomy in the evaluation to understand which types of engaging comments are especially challenging to detect automatically.

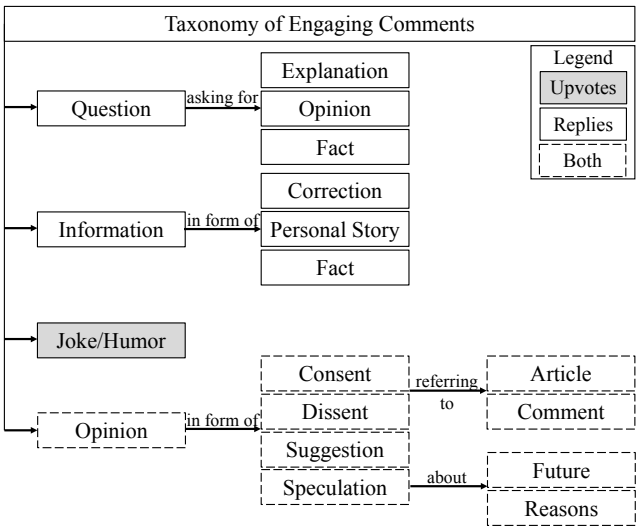


Figure 4: Our hierarchical taxonomy of engaging comments classifies comments that attract upvotes (grey fill), replies (solid outline) or both (dashed outline).

Distinguishing Top and Flop Comments

We present a neural network model to distinguish top and flop comments based on their text. Instead of labeling comments in a time consuming process, we draw upon the number of upvotes and replies that a comment received. To this end, we consider only the comment text and remove the bias of the comment's rank and the news article topic. We build two datasets of top and flop comments, which we then use to train our model with supervised learning.

Removing the Position Bias

We assume that a comment receiving many upvotes or replies is relevant to many users, whereas a comment with no or only a few reactions is comparably irrelevant. However, this assumption only holds if malicious intentions to manipulate the user votes, e.g., voting multiple times with fake user accounts, can be ruled out. In our dataset with 260 million upvotes from *TheGuardian.com*, there is no incentive for users to manipulate the upvote count because it does not influence the order in which comments are displayed. Occasional hoax upvotes can be neglected and considered noise. This leaves us with the vast majority of upvotes actually presenting an engagement signal on the level of individual comments. Still, the number of upvotes and replies is biased by a comment's visibility to readers, which is influenced by the article's popularity and the comment's rank.

News platforms sort comments chronologically and show only the first few (e.g., ten) comments to readers directly below an article text. All subsequent comments are hidden by pagination, which the user can access by browsing to the next page. In practice, most users access only the very first page, which by default shows the oldest comments. They never read any subsequent comments. For this reason, we consider only the first ten comments directly below each article, ensuring that they were seen (and judged) by many readers. Articles with fewer than ten comments are discarded to allow for a fair comparison.

Some news articles draw more attention than others. Thus, they attract a varying number of users who eventually consider voting on and replying to comments. To normalize this variation, we transform the absolute number of upvotes and replies into relative numbers within each article's comment section. To also remove the position bias illustrated in Figure 2, we group all comments across all articles by their rank. The result comprises ten groups of equal size.

We sort the comments of each rank by the descending relative number of upvotes. Each sorted list now contains the comments in a normalized way. All comments in the top 50 percent of the list perform better than an average comment at this rank, which means they received a comparably large portion of upvotes. All comments in the bottom 50 percent received fewer upvotes than an average comment at this rank. Thereby, the list contains top comments and flop comments with regard to upvotes, which can be used as positive

Taxonomy Examples

Question
- **Explanation** "Can anyone explain (serious question!) what the long term economic plan is?"
- **Opinion** "Let's take a poll guesses: What do you think the outcome will be in 17 days..."
- **Fact** "...which [celebrities] are true supporters of Nigel Farage?"

Information
- **Correction** "...is a herbicide, not an insecticide. Please correct this"
- **Personal Story** "The tiered system is incredibly unfair. I have a 16 yr old son, who is extremely..."
- **Fact** "In 2011 ...700,000 new National Insurance numbers were issued to foreign nationals."

Joke/Humor "What is the difference between UKIP and a tandem? A tandem has two seats."

Opinion
- **Dissent (Article)** "I don't like the way this article appears to link Cameron with the polls..."
- **Consent (Comment)** "I agree with [username] that no one can believe a word that he says."
- **Suggestion** "...We need more patriots. We need people that care about the country ..."
- **Speculation**
  - **Future** "...immigration could be $1.5 - -2$ million. Anyone want to argue that will not affect housing,...[?]"
  - **Reasons** "The reason for the Palestinians wanting to have a vatican-like status at the UN is..."

Figure 5: A list of comments exemplifying each class in our hierarchical taxonomy of engaging comments (Figure 4).

and negative training samples for supervised learning.

There is only one variation for processing the replies. Articles that received less than 20 replies on their first ten comments are discarded. In the same way as before, we then sort the comments of each rank by the descending relative number of replies. Splitting each list in halves, results in sets of comments that receive more or fewer replies than an average comment at the respective rank.

By further filtering the dataset, e.g., to only the top 10 percent and bottom 10 percent, we consider only comments that perform much better or much worse than average. This step can be seen as a way to filter for a higher agreement on a comment's rating among users. Typically, upvotes and replies exhibit a low agreement: Users do not agree on which comments deserve upvotes or replies. However, the agreement in the top 10 percent and bottom 10 percent is higher by definition of this subset of the data. A much larger, respectively much lower, relative number of users reacted to the comments in these smaller sets.

Given the positive and negative training samples for supervised learning, we describe the architecture of our neural network model. While we train two separate models, one for upvotes and one for replies as the measure to distinguish top and flop comments, the models have the same architecture. We propose a recurrent neural network model based on Gated Recurrent Units (GRUs) (Cho et al. 2014). The network starts with a pre-trained word embedding layer with fixed weights. We pre-train 300-dimensional word embeddings on our full dataset of 61 million comments. More precisely, we use the skip-gram training method of the fastText algorithm, which allows for mapping even out-of-vocabulary words to embedding vectors (Bojanowski et al. 2017). 4.4 billion tokens are processed, which is about the same number of tokens as in the English Wikipedia. The full text is lowercased and user mentions and URLs are replaced with special tokens. We use the standard size of subwords of 3 to 6 characters and train for 5 epochs. The same pre-trained word embeddings are used for both tasks and thereby the learned word representations are shared across them.

The second layer is a spatial dropout layer, which discards a fraction of the input words for regularization purposes. It is followed by a layer of bidirectional Gated Recurrent Units (GRUs). The bidirectionality allows each unit to consider both previous and subsequent units as context (Schuster and Paliwal 1997). The output of the GRU layer passes a dropout layer and a dense layer. A dense layer with a softmax activation and two outputs handles the final classification. The network is trained with the Adam optimizer and binary cross-entropy as the loss function.

Experiments

The first experiment evaluates the classification accuracy on a dataset of comments from *TheGuardian.com*. We compare four classifiers: (1) logistic regression on text length (baseline); (2) logistic regression on text and user features (Park et al. 2016); (3) a convolutional neural network (CNN) (Kim 2014); and (4) our recurrent neural network based on gated recurrent units (GRU). Second, we use explanation methods for neural networks to investigate which words have the strongest influence on our model's predictions. Finally, we evaluate classification accuracy on another dataset, which consists of product reviews from *Amazon.com*.

User Comments

We consider the task of classifying comments into the classes *top* and *flop 10 percent* with regard to the normalized relative number of upvotes or replies received. For example, a comment classified as *top 10 percent* received a larger relative number of upvotes than 90 percent of the comments with the same rank. We use classification accuracy as the evaluation metric because of the balanced class distribution.

To train the GRU-based model, early stopping on the decrease of validation loss determines the number of training epochs. We set the number of neurons for the GRUs to 32, the dropout to 0.1, and the number of neurons of the dense layer to 16, and refrain from extensive hyperparameter optimization. For comparison, we implement two state-of-the-art approaches: a CNN for sentence classification by Kim (2014) and a feature-based classification approach by Park et al. (2016). Kim's CNN uses a single layer of convolutions and max-over-time pooling. Due to the relatively small number of parameters in this layer, the emphasis is put on the word embedding layer. The feature-based classification approach by Park et al. (2016) was specifically developed to support moderators in identifying high-quality online news comments. It uses the following features: comment length, comment readability, average comment length per user, average comment readability per user, and average number of received comment upvotes per user. These features serve as the input for a logistic regression classifier. In addition to the two approaches from related work, we consider a naive baseline: a logistic regression classifier based solely on comment length.

Results Table 2 shows the accuracy on the task of classifying top and flop comments with regard to upvotes and replies. The column with the name *10* refers to training on a dataset with the two classes *top 10 percent* and *flop 10 percent,* which contains 20k comments. There are two more variants of the experiment also listed in Table 2. The column with the name *25* refers to training on a dataset with the two classes *top 25 percent* and *flop 25 percent,* which contains 53k comments. The column with the name *50* refers to training on a dataset with the two classes *top 50 percent* and *flop 50 percent,* which contains 106k comments. While the training data differ, we use a shared test dataset split from the top/flop 10 percent because the labels in this dataset are the most reliable. The other datasets contain more samples but are noisier. The remaining data for each variant are split into 80 percent training and 20 percent validation. We make sure that there is no overlap between the shared test set and any of the training and validation datasets. Each experiment is repeated ten times.

We perform a paired one-tailed t-test with a 95 percent confidence level to test the significance of our findings. Our null hypothesis is that the true mean difference of the classification accuracy of the GRU and CNN approach is less

Table 2: Classification accuracy on the task of distinguishing top and flop comments on *TheGuardian.com.* with regard to the number of received upvotes and replies.

	Upvotes			Replies		
Top/Flop %	10	25	50	10	25	50
Baseline	0.61	0.61	0.61	0.63	0.63	0.63
Park et al. 2016	0.65	0.66	0.67	0.61	0.59	0.60
Kim 2014	0.67	0.63	0.62	0.69	0.65	0.67
Our Approach	**0.71**	**0.71**	**0.71**	**0.70**	**0.72**	**0.68**

than or equal to zero. The null hypothesis is rejected for all our experiments, leaving us with strong evidence that the GRU approach outperforms the CNN approach with regard to classification accuracy. The results in Table 2 further show that the limitation to the top/flop 10 percent for training in general does not improve classification accuracy. A consequence of this limitation are more reliable labels but also a smaller number of training samples. The GRU approach achieves the best performance on both tasks, upvote and reply prediction. To our surprise, this approach, the logistic regression baseline on comment length only, and the feature-based approach of Park et al. are robust to the different variants of training data (top/flop 10, 25, 50 percent). However, the CNN approach is less robust and performs better if trained on the top/flop 10 percent dataset. If trained on the other dataset variants, the model overfits and does not generalize well to the test data.

Explaining Predictions

Arras et al. (2017) explain neural network predictions in the context of sentiment analysis. We extend their approach to better understand what makes some comments more engaging than others. To this end, we sort all words in the vocabulary according to their relevance for our model predicting *high engagement* in the form of many upvotes or replies. These word relevance scores are calculated with four methods: layer-wise relevance propagation (LRP) (Bach et al. 2015), gradient-based sensitivity analysis (SA) (Li, Monroe, and Jurafsky 2016), integrated gradients (Sundararajan, Taly, and Yan 2017), and a random baseline.

The goal of the experiment is to measure how the deletion of different words changes the classification accuracy of our GRU model. If we consider only true positives, the accuracy in this set is initially 1. The accuracy decreases when we delete the words that are most relevant for the model's prediction and re-run the classification afterward. If we consider only false negatives, the accuracy in this set is initially 0. The accuracy increases when we delete the words that are least relevant for the correct class and re-run the classification. The words that are deleted speak against the correct class. Therefore, if their deletion changes the classification in favor of the correct class, accuracy increases.

Figure 6 visualizes how deleting the most/least relevant words affects classification accuracy of our GRU model. The larger the change in accuracy, the better are the calculated word relevance scores. The two methods LRP and integrated

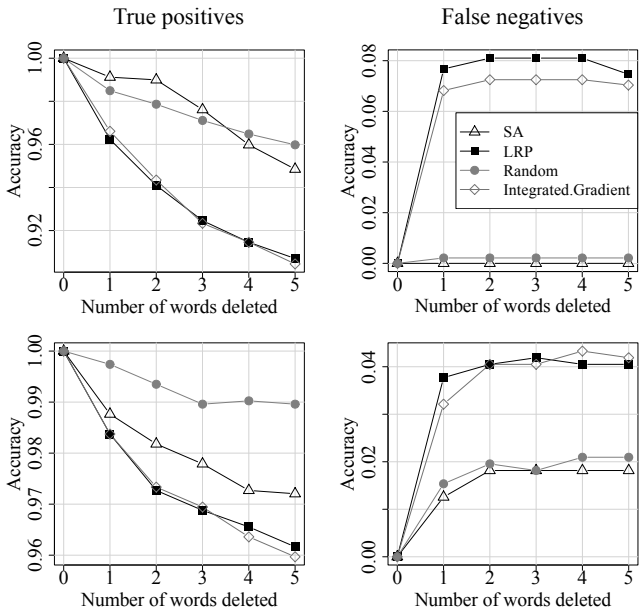


Figure 6: Deleting the most relevant words from true positives (left-hand part) and the least relevant words from false negatives (right-hand part) has the strongest effect on reply (upper part) and upvote prediction (lower part) when using LRP relevance scores.

gradients provide almost the same relevance scores and both outperform the method SA and the random baseline.

Based on layer-wise relevance propagation (LRP) (Bach et al. 2015), we identify the most and least relevant words for our model's decisions. Words that refer to strong emotions or controversial topics *(arrogant, depressing, fantastic, bearable, Brexit)* are most relevant for predicting upvotes. Least relevant are stop words *(won't, wasn't)* or emotions that are typically expressed in short comments *(lol, sigh)*. Most relevant for predicting many replies are words referring to the Labour party *(socialist, lefty)*, which corresponds to the political orientation of most *TheGuardian.com* readers. The least relevant words are names of British public figures *(Pickles, Keir, Tanner, Morgan)*.

According to our taxonomy for engaging comments, we labeled all positive samples in the test set of the *top/flop 10 percent* dataset. For each class, Figure 7 shows our model's recall at distinguishing top and flop comments. The classes *Correction* and *Comment Consent* are omitted because there was only a handful of such samples in the test set. The recall for *Joke/Humor* is lowest, whereas the recall for *Comment Dissent* or speculation about *Future* and *Reasons* is highest. This discrepancy means that our model's predictions could be improved by a better detection of *Joke/Humor*. Further, questions asking for facts (*Q:Fact*) are identified with higher recall than comments providing facts (*Fact*). Besides these differences, the recall for all classes is similar.

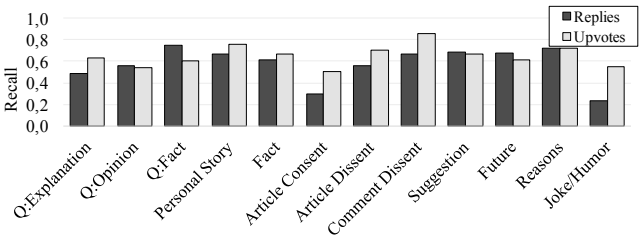


Figure 7: Recall for identifying engaging comments differs per class, e.g., jokes are less reliably identified than dissent.

Product Reviews

User comments on news platforms and product reviews on online retail platforms have several properties in common: (1) popular news articles, as well as popular products, generate an overwhelming number of posts, (2) posts on both platforms are typically short, and (3) both allow users to vote on posts. On product review platforms, upvotes resemble votes on the helpfulness of a review. However, news discussions differ from disconnected posts on Amazon, YouTube, and Twitter, where no discussions take place and the communication is unidirectional. Reviews focus on a particular product and do not refer to each other. Danescu-Niculescu-Mizil et al. (2009) analyze a dataset of Amazon product reviews and their helpfulness votes. They find that users consider a review more helpful if the associated product rating is closer to the average rating for this product (conformity bias).

We consider product reviews posted on *Amazon.com* to study the applicability of our approach to other domains. The dataset contains 82 million Amazon product reviews, spanning from May 1996 to July 2014, and is available online (He and McAuley 2016). 170 million upvotes ("Was this review helpful?") were cast in total. We filter the dataset so that we consider the ten earliest reviews per product. Products with less than ten reviews are discarded. Similar to the dataset of user comments at *TheGuardian.com*, we learn word embeddings on this large dataset. The reviews comprise 7.6 billion tokens, which is more than twice the number of tokens in the English Wikipedia.

For a classification experiment, we use a subset of 9 million book reviews to reduce topical variety. We apply the same normalization steps to upvote counts as described earlier and consider three different variants of the dataset. They correspond to the top and flop 10, 25, and 50 percent of the product reviews and contain 220k, 550k, and 1.1 million product reviews, respectively. The test set is shared for all variations of the training data and comprises 10 percent of the *top/flop 10 percent* dataset (22k reviews). The remaining data for each variant are randomly split into 80 percent training and 20 percent validation set.

We compare the classification accuracy of the logistic regression baseline on review length, the CNN by Kim (2014), and our approach on the task of distinguishing helpful (top) and non-helpful (flop) product reviews. The feature-based classifier by Park et al. (2016) cannot be applied to the product reviews because it requires user information, which our dataset does not contain.

Table 3: Classification accuracy on the task of distinguishing top and flop product reviews on *Amazon.com* with regard to the number of received helpfulness upvotes.

Top/Flop Percent	10	25	50
Baseline	0.67	0.67	0.34
Kim 2014	0.67	0.72	0.64
Our Approach	**0.76**	**0.75**	**0.66**

Results Table 3 lists the results of the experiment. The GRU model outperforms the CNN. In contrast to our comment dataset, the limitation to the top/flop 10 percent on the product reviews dataset improves classification accuracy. Here, the training dataset is ten times larger, which diminishes the disadvantage of limiting the data to the top and flop 10 percent. The more reliable labels in the top/flop 10 and 20 percent training datasets make the difference.

Training on the top/flop 50 percent dataset results in the worst performance. For the baseline that considers only the comment length, it is even worse than random guessing, which achieves 50 percent accuracy. The different value distributions of review lengths in training and test data explains this result. The baseline is unable to learn an appropriate threshold for the comment length. The most and least engaging product reviews in the top/flop 50 percent dataset have similar average length (1076 vs. 1055 characters), whereas there is a clear separation for review lengths in the top/flop 10 percent dataset (677 vs. 1387 characters).

Impact on Online Discussions

Many online news platforms closed their comment section under the unbearable workload of content moderation and hateful and abusive comments. But, hateful comments and abusive language are not the only problem. Without in-depth discussions, where users exchange reasonable arguments for their opinions, comment sections create (almost) no added value for the news platforms. Users shout out their own opinions but rarely ever listen to each other and start fruitful conversations. To stand out among the plethora of online spaces where users can post their opinions, modern news platforms need to add value by providing a space for engaging and polite exchange. Any change to comment sections that fosters user interaction or increases commitment by design (Aragón, Gómez, and Kaltenbrunner 2017; Farzan et al. 2011; Budak et al. 2017) can make a big difference.

Today, comments in online discussions are mostly ranked chronologically — with a few exceptions, such as *Slashdot.org* and *Digg.com*. While one might argue that this approach is transparent and fair, it does not foster engaging discussions. Instead, it only gives an incentive to post comments as fast as possible after an article is published. In that case, the comment will get ranked high, it will gain visibility in the community, and possibly get some reactions to the comment, no matter how good or bad it is. This competition goes so far that some users refrain from reading the article to be the first to post a comment. Our approach introduces an

alternative method to rank comments by the expected number of upvotes and replies. If applied, on the one hand, the visibility of top-performing, engaging comments increases. They are shown to more users. On the other hand, the least engaging, flop comments lose visibility and are practically hidden at the end of the comment section, which usually no user accesses.

A limitation of our study is that we only consider a comment's text content and no user-based features. The reputation of the comment author presumably affects its impact in terms of visibility and thus received upvotes and replies. Further, the most comment texts that we explored are well-formed and grammatically correct, which simplifies the analysis. Emoticons and slang are rarely used on *TheGuardian.com*. However, they might be more frequent on other platforms and pose a potential challenge. Design changes on the online platforms are an additional challenge for analyzing a long timespan. For example, with the most recent features, users can sort comments by time or by the number of upvotes. The default setting of the sorting, e.g. newest/oldest first, is an important factor for the visibility of individual comments. Editor picks change the visibility of selected comments in a similar way.

Conclusions and Future Work

We studied comment upvotes and replies as a measure of user engagement in online news discussions. To this end, we designed a taxonomy of engaging comments on the platform *TheGuardian.com* and analyzed textual differences of the most and least engaging comments (top and flop comments). Further, we trained a neural network model to distinguish these top and flop comments given the comments' texts. To construct the training dataset, we identified and removed the position bias that favors early comments and normalized upvote counts for each article individually to also remove a potential topical bias.

Based on predicted user reactions in the form of upvotes and replies, platforms could automatically highlight or rank comments to show top ones to users and thus encourage more interaction. Experimental results demonstrate that neural network models outperform feature-based classification approaches and achieve an accuracy of about 70 percent on a balanced test dataset. This result is not limited to our dataset of comments and also generalizes to product reviews.

A promising path for future work is to investigate different types of votes, e.g., not only upvotes but also downvotes or more fine-grained votes. It would be interesting to analyze the interplay of these types. A comment that receives many upvotes and downvotes at the same time might be considered controversial. Another idea is to consider user names and reputation as a predictive feature in the classification process. For example, a comment by a journalist might generate more engagement than a comment by a regular user.

Acknowledgments

We would like to thank Johannes Filter, Cornelius Hagmeister, and Thomas Kellermeier for their contribution to this project during our seminar *Text Mining in Practice*.

References

Aldous, K. K.; An, J.; and Jansen, B. J. 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)*, volume 13, 47–57.

Ambroselli, C.; Risch, J.; Krestel, R.; and Loos, A. 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)*, 193–199.

Aragón, P.; Gómez, V.; García, D.; and Kaltenbrunner, A. 2017. Generative models of online discussion threads: state of the art and research challenges. *Journal of Internet Services and Applications* 8(1):15.

Aragón, P.; Gómez, V.; and Kaltenbrunner, A. 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)*, 12–21.

Arras, L.; Montavon, G.; Müller, K.-R.; and Samek, W. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 159–168.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10(7):1–46.

Backstrom, L.; Kleinberg, J.; Lee, L.; and Danescu-Niculescu-Mizil, C. 2013. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, 13–22.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the International Conference on World Wide Web Companion (WWW)*, 759–760.

Belam, M. 2011. Adding responses to comments. https://www.theguardian.com/help/insideguardian/2011/nov/03/responses-in-comments. *The Guardian.* Accessed: 2020-03-26.

Berry, G., and Taylor, S. J. 2017. Discussion quality diffuses in the digital public square. In *Proceedings of the International Conference on World Wide Web (WWW)*, 1371–1380.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Budak, C.; Garrett, R. K.; Resnick, P.; and Kamin, J. 2017. Threading is sticky: How threaded conversations promote comment system user retention. *Proceedings of the ACM on Human-Computer Interaction (HCI)* 1(CSCW):1–20.

Chen, H.; Mckeever, S.; and Delany, S. J. 2017. Presenting a labelled dataset for real-time detection of abusive user posts. In *Proceedings of the International Conference on Web Intelligence (WI)*, 884–890.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 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)*, 1724–1734.

Curtis, C. 2017. How britain voted at the 2017 general election. https://yougov.co.uk/topics/politics/articles-reports/2017/06/13/how-britain-voted-2017-general-election. *YouGov.* Accessed: 2020-03-26.

Danescu-Niculescu-Mizil, C.; Kossinets, G.; Kleinberg, J.; and Lee, L. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the International Conference on World Wide Web (WWW)*, 141–150.

Diakopoulos, N., and Naaman, M. 2011. Towards quality discourse in online news comments. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*, 133–142.

Diakopoulos, N. 2015. Picking the nyt picks: Editorial criteria and automation in the curation of online news comments. *Journal of the International Symposium on Online Journalism (ISOJ)* 6(1):147–166.

Farzan, R.; Dabbish, L. A.; Kraut, R. E.; and Postmes, T. 2011. Increasing commitment to online communities by designing for social presence. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*, 321–330.

Gómez, V.; Kappen, H. J.; Litvak, N.; and Kaltenbrunner, A. 2013. A likelihood-based framework for the analysis of discussion threads. *World Wide Web* 16(5-6):645–675.

Gómez, V.; Kaltenbrunner, A.; and López, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the International Conference on World Wide Web (WWW)*, 645–654.

Hanman, N. 2012. Threading arrives on comment is free. https://www.theguardian.com/commentisfree/2012/dec/03/threading-arrives-on-comment-is-free. *The Guardian.* Accessed: 2020-03-26.

He, R., and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the International Conference on World Wide Web (WWW)*, 507–517.

Hsu, C.-F.; Khabiri, E.; and Caverlee, J. 2009. Ranking comments on the social web. In *Proceedings of the International Conference on Computational Science and Engineering (CSE)*, 90–97.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Kobayashi, R., and Lambiotte, R. 2016. TiDeH: Time-dependent hawkes process for predicting retweet dynamics. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*.

Kolhatkar, V., and Taboada, M. 2017. Using new york times picks to identify constructive comments. In *Proceedings of the Natural Language Processing meets Journalism Workshop (NLPmJ@EMNLP)*, 100–105.

Kumar, R.; Mahdian, M.; and McGlohon, M. 2010. Dynamics of conversations. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 553–562.

Lakkaraju, H.; McAuley, J.; and Leskovec, J. 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)*.

Lampe, C., and Resnick, P. 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)*, 543–550.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* 1–18.

Medvedev, A. N.; Delvenne, J.-C.; and Lambiotte, R. 2019. Modelling structure and predicting dynamics of discussion threads in online boards. *Journal of Complex Networks* 7(1):67–82.

Napoles, C.; Tetreault, J.; Pappu, A.; Rosato, E.; and Provenzale, B. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, 13–23.

Napoles, C.; Pappu, A.; and Tetreault, J. R. 2017. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 628–631.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *Proceedings of the International Conference on World Wide Web (WWW)*, 145–153.

Park, D.; Sachar, S.; Diakopoulos, N.; and Elmqvist, N. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 1114–1125.

Risch, J., and Krestel, R. 2018. Delete or not delete? semi-automatic comment moderation for the newsroom. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*, 166–176.

Rizos, G.; Papadopoulos, S.; and Kompatsiaris, Y. 2016. Predicting news popularity by mining online discussions. In *Proceedings of the International Conference on World Wide Web Companion (WWW)*, 737–742.

Salminen, J.; Almerekhi, H.; Milenković, M.; Jung, S.-g.; An, J.; Kwak, H.; and Jansen, B. J. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 330–339.

Schabus, D., and Skowron, M. 2018. Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 1602–1605.

Schabus, D.; Skowron, M.; and Trapp, M. 2017. One million posts: A data set of german online discussions. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, 1241–1244.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Siersdorfer, S.; Chelaru, S.; Nejdl, W.; and San Pedro, J. 2010. How useful are your comments? analyzing and predicting youtube comments and comment ratings. In *Proceedings of the International Conference on World Wide Web (WWW)*, 891–900.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 3319–3328.

Wang, C.; Ye, M.; and Huberman, B. A. 2012. From user comments to on-line conversations. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 244–252.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 78–84.

Wei, Z.; Liu, Y.; and Li, Y. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, 195–200.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the International Conference on World Wide Web (WWW)*, 1391–1399.

Xu, J.-M.; Jun, K.-S.; Zhu, X.; and Bellmore, A. 2012. Learning from bullying traces in social media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 656–666.

Zhang, Q.; Gong, Y.; Wu, J.; Huang, H.; and Huang, X. 2016. Retweet prediction with attention-based deep neural network. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, 75–84.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1350–1361.