Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus

Roxanne El Baff **Henning Wachsmuth** * **Khalid Al-Khatib** **Benno Stein**
Bauhaus-Universität Weimar, Weimar, Germany, <first>{.<last>}+@uni-weimar.de
* Paderborn University, Paderborn, Germany, henningw@upb.de

Abstract

News editorials are said to shape public opinion, which makes them a powerful tool and an important source of political argumentation. However, rarely do editorials change anyone's stance on an issue completely, nor do they tend to argue explicitly (but rather follow a subtle rhetorical strategy). So, what does argumentation quality mean for editorials then? We develop the notion that an effective editorial challenges readers with opposing stance, and at the same time empowers the arguing skills of readers that share the editorial's stance — or even challenges both sides. To study argumentation quality based on this notion, we introduce a new corpus with 1000 editorials from the New York Times, annotated for their perceived effect along with the annotators' political orientations. Analyzing the corpus, we find that annotators with different orientation disagree on the effect significantly. While only 1% of all editorials changed anyone's stance, more than 5% meet our notion. We conclude that our corpus serves as a suitable resource for studying the argumentation quality of news editorials.

1 Introduction

A news editorial is an article that argues in favor of a particular stance on a usually timely controversial issue, such as the relocation of the US embassy in Israel to Jerusalem. Usually, it reflects the political ideology of the newspaper, aiming to persuade readers of the respective stance. Such editorials are said to have the power to shape the opinion of the masses. Similarly, they can increase or decrease the gap between readers with opposing beliefs (van Dijk, 1995). As such, news editorials represent an important resource for research on argument mining (Mochales and Moens, 2011) and debating technologies (Rinott et al., 2015).

On the other hand, a single news editorial rarely changes the stance of a reader completely. Moreover, many editorials do not put an explicit focus on arguments. Rather, they follow a subtle rhetorical strategy combining emotional anecdotes with hidden claims and ethotic evidence, among others (Al-Khatib et al., 2017). So, if not persuasive arguments, what makes a news editorial effective or ineffective then? In other words: How can we measure its argumentation quality?

In this paper, we introduce a new corpus with 1000 news editorials from the New York Times where we consider argumentation quality from a dialectical perspective (van Eemeren and Grootendorst, 2004). While several quality dimensions are known in theory (Wachsmuth et al., 2017b), existing approaches rely on subjective assessments of absolute (Persing and Ng, 2015) or relative (Habernal and Gurevych, 2016) persuasiveness. In contrast, our corpus captures quality in terms of whether an editorial brings readers of opposing belief closer together or rather increases the gap between them. We argue that, thereby, we better account for the practically achieved persuasive effect, resulting in a qualitative media measurement analysis of editorials that intrigue our thoughts.

Persuasion, according to Halmari and Virtanen (2005), is an umbrella term for linguistic choices that aim at changing or affecting the behavior of others or at strengthening the existing beliefs of those who already agree, including the persuaders themselves. To study persuasion for editorials, four dimensions must be considered: (1) prior beliefs of readers, (2) prior beliefs and behaviors of authors, (3) effects of the text, and (4) linguistic choices. We account for these dimensions as follows.

Prior Beliefs of Readers Given the focus of news editorials on timely politics, we use the political typology quiz[1] developed by the Pew Research

[1]Typology quiz, http://www.people-press.org/quiz/political-typology/

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 454–464
Brussels, Belgium, October 31 - November 1, 2018.

Center to measure the prior beliefs of readers. The underlying typology divides Americans into eight political groups, as detailed later on: four largely liberal and four largely conservative ones, along with a ninth group of the politically less engaged.

Prior Beliefs and Behaviors of Authors Each newspaper has its set of beliefs, reflected in particular stances on different controversial issues. To avoid newspaper-related side effects in the study of argumentation quality, we decided to control this dimension by annotating news editorials from one source only. In particular, we resort to the online portal of the New York Times for two practical reasons: (1) The political typology quiz is tailored to people from the United States. (2) A large source of news editorials and detailed metadata is already available (Sandhaus, 2008).

Effects of the Persuasive Text We tackle the outlined dialectical view of argumentation quality by asking annotators about how a given news editorial affected them: If you have a different stance than the editorial, did it *challenge* you, making you rethink your stance? Or, if you have the same stance, did it *empower* you, enabling you to better defend your stance? We postulate that high-quality argumentation challenges one side and empowers the other side at the same time, and we hypothesize that this notion allows distinguishing effective and ineffective editorials regardless of the annotators' stance. We analyze the corpus in order to investigate this hypothesis in comparison to classical approaches asking for persuasion.

Linguistic Choices Our goal is to provide a resource for studying the quality of editorial argumentation and their underlying rhetorical strategies. Accordingly, we leave an analysis of the linguistic features impacting quality to future research.

The contribution of this paper is three-fold:

- We propose a new notion of argumentation quality for news editorials based on how challenging and empowering an editorial is for readers with opposing stances.
- We create a freely available corpus[2] with quality assessments of 1000 news editorials, each annotated by three liberals and three conservatives. The annotators also reported free-text reasons for the effects they observed.
- We analyze the corpus, finding that more than 5% of all editorials fulfill our notion of high quality, whereas only 1% really persuaded any annotator. As expected, annotators agree only when sharing similar prior beliefs.

[2]Webis-Editorial-Quality-18 corpus, available at http://www.webis.de/data

2 Related Work

Computational argumentation has lately become popular in the natural language processing community. So far, most computational argumentation research deals with the mining of arguments from text (Mochales and Moens, 2011). Accordingly, many studied corpora capture argument structure, often for a specific text genre, such as persuasive essays (Stab and Gurevych, 2014), Wikipedia articles (Levy et al., 2014), or even pure arguments (Peldszus and Stede, 2015). These genres share that they make claims and reasons explicit, i.e., they argue rationally. In contrast, real-world argumentation related to politics often comprises more sophisticated mechanisms, bringing together logical arguments (Johnson and Blair, 2006) with rhetorical means (Aristotle, translated 2007) and dialectic (van Eemeren and Grootendorst, 2004). A typical genre of such kind is *news editorials*.

As outlined in Section 1, news editorials are opinionated articles that aim to persuade their readers of a stance towards some controversial issue, usually with implicit, hidden strategies (van Dijk, 1995). Editorials have been used for opinion mining and retrieval in some works (Yu and Hatzivassiloglou, 2003; Bal, 2009), partly towards analyzing argumentation (Bal and Dizier, 2010; Kiesel et al., 2015). To our knowledge, the only corpus of noteworthy size that exists for studying editorial argumentation explicitly is the one of Al-Khatib et al. (2016) who segmented 300 editorials into argumentative discourse units of different claim and evidence types.

Al-Khatib et al. (2017) trained classifiers on their corpus and applied them to 28,986 editorials from the New York Times Annotated Corpus (Sandhaus, 2008). They found topic-specific evidence type patterns, which appear to be related to persuasive strategies. However, editorial-level annotations are missing that actually connect the patterns to persuasiveness. For blog posts and forum discussions respectively, previous work has annotated persuasive acts (Anand et al., 2011) and the use of Aristotle's rhetorical means (Hidey et al., 2017). Still, this would not allow distinguishing effective from

ineffective strategies. We fill this gap by presenting the first editorial corpus with persuasion-related annotations of *argumentation quality*. To obtain a larger corpus size, we rely on the editorials from Sandhaus (2008) rather than those from Al-Khatib et al. (2016). Potash et al. (2017) observe bias in existing corpora towards higher quality for longer arguments. To prevent such bias, we consider only editorials from a narrow length range.

Research on argumentation quality has recently been surveyed by Wachsmuth et al. (2017b). The authors developed a taxonomy with one main aspect each for logical (*cogency*), rhetorical (*effectiveness*), and dialectical quality (*reasonableness*), as well as several concrete quality dimensions. Effectiveness reflects to what extent an author persuades a reader, and reasonableness reflects an argument's contribution to agreement. As detailed in Section 3, the dimension we propose is meant to measure persuasive effectiveness, yet, from a dialectical perspective, which is more suitable for editorials. We hypothesize it to be related to the acceptability of arguments (Cabrio and Villata, 2012) and the helpfulness of argumentation (Liu et al., 2017).

While Louis and Nenkova (2013) study the general quality of news articles, our goal is to provide a basis for studying their argumentation quality more objectively. Some existing computational approaches to assessing argumentation quality rely on human persuasiveness ratings of essays (Persing and Ng, 2015; Wachsmuth et al., 2016) or debate portal arguments (Persing and Ng, 2017). The problem here is that persuasiveness is subjective by heart, underlined by the low inter-annotator agreement for effectiveness in the corpus of Wachsmuth et al. (2017b): effectiveness depends on the prior stance of the annotator.

Habernal and Gurevych (2016) compare the convincingness of arguments with only one stance on a given issue, which circumvents the problem, but does not help for actual persuasion. While Tan et al. (2016) analyze how people are persuaded by others with opposing stance, they restrict their view to good-faith discussions (where people are open to be persuaded) — a setting not common for political argumentation. Instead, we tackle subjectiveness by letting people with both stances on a discussed issue annotate quality.

Cano-Basave and He (2016) point out that persuasive argumentation is about both changing and reinforcing stance — a view that we follow. The authors study the impact of persuasive language of political debates based on poll changes. Such a direct effect on different audiences is not accessible for most argumentative texts, including editorials. Persuasiveness does not only depend on a text itself, but also on the reader's beliefs and personality. Lukin et al. (2017) find that different types of arguments (rational vs. emotional) are effective depending on the "Big Five" personality traits (Goldberg, 1990). We captured our annotators' personality traits, too. However, we primarily focus on nine political profiles from left to right (Doherty et al., 2017) in order to represent prior stance. We are not aware of any previous work in computational argumentation considering such profiles so far.

3 A New Model of Argumentation Quality for News Editorials

We propose a model that quantifies the argumentation quality of an editorial at the discourse level. Two dimensions of persuasion are considered to this end: the *prior beliefs of the reader* and the *effect of the text*. Regarding the former, readers are profiled based on a political typology. Regarding the latter, we annotate an editorial's capability to either (1) challenge or to (2) reinforce a reader's stance; we also consider the magnitude of the effect. Based on the annotations, for which we consider a set of annotators belonging to at least two spectrums of beliefs (three annotators for each), the quality of an editorial can be assessed.

3.1 Prior Beliefs of a Reader

The existing beliefs of the reader of an editorial are a crucial factor when measuring the editorial's argumentation quality. Theoretically, it would be best to consider two reader groups for each editorial that have an opposite stance on the concrete issue discussed in the editorial. Practically, finding such readers for a considerable number of editorials is hardly feasible, because the reader's stance on the issue is not accessible beforehand.

As a proxy, we therefore decided to model the reader's prior beliefs by identifying the reader's political ideology. In particular, we profile the reader as being *liberal* or *conservative* based on the nine groups of the political typology developed by the PEW Research Center. The typology includes four groups that belong to the liberal ideology and four that belong to the conservative ideology:[3]

[3]The ninth rather small group is the *bystanders*, which we

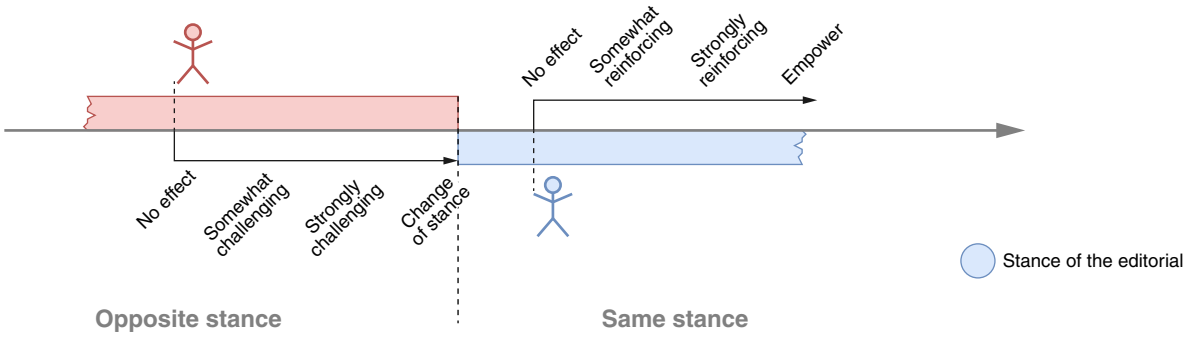

Figure 1: Illustration of potential effects of a news editorial on readers from two belief groups: Those whose prior stance matches the stance of the editorial on the discussed issue, and those whose prior stance opposes it.

Liberal Ideologies *solid liberals*, *opportunity democrats*, *disaffected democrats*, and *devout and diverse*.

Conservative Ideologies *core conservatives*, *country first conservatives*, *market skeptic republicans*, and *new era enterprisers*.

3.2 Effect of the Text

We measure the effect of a news editorial along two characteristics: how *challenging* and how *reinforcing* the editorial is. An editorial is challenging, if it makes the reader rethink his or her prior stance on the discussed issue, even though he or she may not change the stance in the end. On the other hand, an editorial is reinforcing, if it helps the reader in building or further corroborating his or her prior stance on the discussed topic. To capture the magnitude of the effect of the editorial's text, we consider the following labels:

- *Strongly challenging:* The editorial made the reader really rethink whether and why he/she thinks that his/her prior stance is right.
- *Somewhat challenging:* The editorial conveyed at least some information opposite to the reader's stance that was new and noteworthy for him/her.
- *No effect:* The reader did not find any new and noteworthy information opposing or supporting his/her prior stance.
- *Somewhat reinforcing:* The editorial conveyed at least some information supporting the reader's stance that was new and noteworthy for him/her.
- *Strongly reinforcing:* The editorial enabled the reader to argue really better for his/her stance.

The ultimate goal of a news editorial is to *change the stance* of readers with an opposite prior stance. An editorial may reach this effect in case it strongly challenges the reader. On the other hand, in case a reader already has the same stance as the editorial, then the ultimate goal of a news editorial is to *empower the reader* to argue better for his or her stance on the discussed issue. Analog to the previous case, an editorial may have this effect in case it strongly reinforces the stance of the reader. The potential effects on readers from opposing belief groups are visually illustrated in Figure 1.

3.3 Editorial Argumentation Quality

We argue that the argumentation quality of a news editorial is governed by two factors: (1) whether the news editorial increases or decreases the gap between readers with opposing beliefs (van Dijk, 1995) and (2) whether the news editorial presents new and/or persuasive argumentation.

Having the effect labels assigned by readers of opposing belief groups A and B, we follow the dialectical perspective outlined in Section 1 to interpret the argumentation quality of each possible combination *A & B* of labels as follows:

- *Challenging & Challenging:* The editorial challenges the stance of both groups. This suggests that it comprises new and noteworthy argumentation for understanding each other's stance. We see this as an indicator of high quality, since it helps bringing the two groups closer together.
- *Challenging & Reinforcing:* The editorial challenges the stance of one group and reinforces the stance of the other. This suggests

ignore, since it represents those people that are considered not involved in what is happening in politics. About 8% of the American population are supposed to be bystanders.

		Group A		
		Challenging	*Reinforcing*	*No Effect*
Group B	*Challenging*	**High quality.** Brings groups closer together. New and persuasive argumentation.	**High quality.** Helps agreeing on one stance. New and persuasive argumentation.	**Medium quality.** Helps agreeing on one stance. Persuasive argumentation.
	Reinforcing		**Medium quality.** New argumentation.	**Rather low quality.** Increase the gap between the groups. New argumentation.
	No Effect			**Low quality.** Neither new nor persuasive.

Table 1: Interpretation of the combined effects and quality of a news editorial for two groups with opposing beliefs.

that it comprises new and persuasive argumentation in favor of one stance. We see this as an indicator of high quality, too, since it does not only help the two groups agree on the same stance, but also further supports that stance.

- *Challenging & No Effect:* The editorial challenges the stance of one group but does not affect the other. This suggests that it comprises persuasive but not fully new argumentation in favor of one stance. We see this as an indicator of medium quality, since it at least helps the two groups agree on the same stance.
- *Reinforcing & Reinforcing:* The editorial reinforces the stance of both groups. This suggests that it comprises new and noteworthy argumentation for clarifying the two possible stances. We see this as an indicator of medium quality, since it at least provides new insights into the discussed issue.
- *Reinforcing & No Effect:* The editorial reinforces the stance of one group but does not affect the other. This suggests that it comprises new but not fully persuasive argumentation in favor of one stance. We see this as an indicator of rather low quality, since it increases the gap between the two groups.
- *No Effect & No Effect:* The news editorial does not affect either group. This suggests that it comprises neither new nor persuasive argumentation. We see this as an indicator of low quality, since it makes the need for the editorial questionable.

Table 1 summarizes our interpretation of the effects and their quality for each combination.

4 Corpus Construction

Based on the model from Section 3, we conducted an annotation study to build a new corpus for studying the argumentation quality of news editorials. This section describes how editorials were acquired, sampled, and annotated. Furthermore, it presents an overview of the resulting corpus.

4.1 Editorial Acquisition and Sampling

As mentioned before, we decided to restrict our study to editorials from a single news portal (The New York Times), in order to exclude the portal impact on the quality assessment. Particularly, we used the New York Times Annotated Corpus (Sandhaus, 2008), which comprises around 1.8 million news articles written between 1987 and 2007. Each of these articles comes with 27 metadata tags capturing the article's type, topic, author, etc.

To identify news editorials, we used the tags *descriptor* and *taxonomic classifiers* with the values 'Opinion' and 'Editorial'. To maximize recency, we considered those written between 2005 and 2007 only. This resulted in 2556 editorials with a mean length of 492 words. To control the length, we left out short editorials (< 450 words) and long ones (> 650 words), ending up with 1022 editorials. We randomly selected five of these for the pilot study and 1000 for the main one.

4.2 Editorial Annotations

Carrying out the task of annotating all editorials in our corpus was divided into three phases: (1) the selection of annotators, (2) a pilot study, and (3) the main annotation. After discussing the annotation task, we explain the three phases in detail.

Annotation Task As shown in Table 2, we asked our annotators to assess the effect of each editorial's content along the five labels from Section 3 as well as a potential empowering or change of stance. Given an editorial, an annotator should first read its text carefully and then answer the question 'How did the news editorial affect you?' (question

#	Questions	Answers
1	How did the news editorial affect you?	a. It strongly challenged my stance b. It somewhat challenged my stance c. It neither challenged nor reinforced my stance d. It somewhat reinforced my stance e. It strongly reinforced my stance
1a	Did the editorial actually change your stance on the discussed issue (from pro to con, or vice versa)?	Yes / No
1e	Did the editorial empower you to better argue for your stance?	Yes / No
2	Explain your choice(s) (Keep it short)	*Free text*

Table 2: The questions that our annotators had to answer after reading each news editorial. Only in case option *a* was chosen for question 1, question 1a was asked. Accordingly, only in case option *e* was chosen for question 1, question 1e was asked. In any case, the annotator was asked to explain his or her answers (question 2).

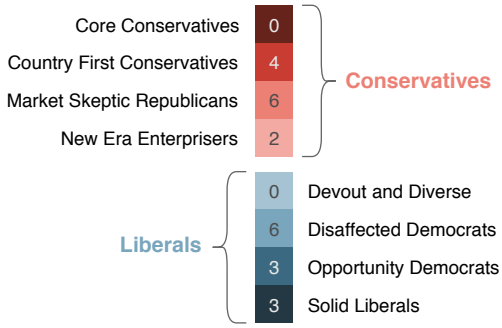

Figure 2: The distribution of the 24 selected annotators over the eight considered political ideologies.

1). The possible answers ranged from *strongly challenging* to *strongly reinforcing*. In case the answer was *strongly challenging*, we asked whether it actually changed his or her stance (question 1a). In case the answer was *strongly reinforcing*, we asked whether it actually empowered him or her to argue better about the topic (question 1e). Finally, the annotator should briefly explain the rationale of his or her choice(s) (question 2).

Selection of Annotators We recruited native English speakers from the United States with at least a bachelor's degree via *upwork.com*. We asked them to do the PEW political typology quiz and to report their results. 40 candidates were recruited until we had both 12 annotators with liberal ideology and 12 with conservative ideology. Figure 2 illustrates the distribution of annotators over the possible political ideologies. The annotators within each group, liberals or conservatives, were selected randomly.

Pilot study We conducted a pilot study with five randomly chosen news editorials in order to check whether our annotation guidelines were clear, consistent, and understood by the annotators. We randomly selected three liberals (one *solid liberal* and two *disaffected democrats*) and three conservatives (one *core conservative*, *market skeptic republican*, and *new era enterpriser* each) to annotate all five editorials. We solicited the annotators to give a feedback about the guidelines and the process. All six annotators reported that the guidelines were easy to follow and easy to understand. Also, they gave some suggestions which we followed by rephrasing some questions in the guidelines.

Main Annotation We divided the sampled 1000 news editorials into four batches of size 250. Each batch was assigned to six annotators based on their political ideology (three liberals and three conservatives). As a result, we obtained 6000 different annotations. To prevent annotators from prejudging an editorial, we kept the source of the editorials as well as their titles hidden. Thereby, we focused on the editorial's content while leaving a study of the impact of its title to future work. As mentioned before, the editorials were somewhat outdated, hence, we asked the annotators to try to think back to when the discussed issue was current.

The annotation was done using a web application that we developed specifically for this purpose. Each annotator had to login with an assigned identification number and his or her result of the political typology quiz.

4.3 Corpus Overview

Table 3 shows statistics of the resulting corpus. The most frequently annotated effect was *somewhat reinforcing* followed by *no effect*. The rarest in turn was *strongly challenging* for both liberals and conservatives with only 143 out of 6000 annotations (i.e., 2.4%). Even more rare, in only 68 cases the reader actually changed his or her stance which is equivalent to about 1% of all annotations.

Political Orientation	Effect with Intensity: Strongly challenging (change)	Somewhat challenging	No effect	Somewhat reinforcing	Strongly reinforcing (empower)	Effect without Intensity: Challenging	No effect	Reinforcing
Liberals	71 (33)	269	708	1402	550 (509)	340	708	1952
Conservatives	72 (35)	275	1282	798	573 (461)	347	1282	1371
Overall	**143 (68)**	**544**	**1990**	**2200**	**1123 (970)**	**687**	**1990**	**3323**

Table 3: Counts of the annotated effects for the 1000 news editorials in our corpus depending on the annotators' political orientation, once with intensity (strongly vs. somewhat), once without. In parentheses: Annotators that changed stance or felt empowered. Each editorial was annotated by three liberal and three conservative annotators.

[...] Police officers firing 50 rounds early last Saturday killed Sean Bell, an unarmed man who was to have married his high school sweetheart later in the day. The mayor and the commissioner moved quickly to answer questions and to hear the concerns of the victim's family and the community. But their responsiveness will not bring back Sean Bell. The challenge here is far greater than good communications. The officers who killed Mr. Bell were part of a sting operation at a Queens nightclub suspected of narcotics, prostitution and weapons violations. According to published reports, the officers have said that as Mr. Bell and his friends left the club and headed toward their car, an undercover detective heard one of them say he was going to get a gun. They also reportedly said that when the men entered the car, the detective pulled his gun and identified himself, but the car suddenly gunned forward, hit him in the shin and then struck an unmarked police minivan. The officers then opened fire. The tragedy may simply involve two sets of very frightened men who reacted instinctively to what they thought was imminent danger. But only one of the sets was armed. There was no gun in the car, nor on the shooting victims, who sat helpless inside while five police officers began firing 50 rounds at them. [...]

Figure 3: An excerpt of the news editorial "50 Bullets and a Death in Queens". This editorial challenged the stance of annotators with conservative ideology and reinforced the stance of those with liberal ideology.

Weather is not primarily a moral affair. We do not deserve a long, slow patch of hot weather, like the one that sat on the city in early June, any more than we deserve the extraordinarily beautiful evenings that have come with these longest days of the year. Deserving has nothing to do with it. The weather comes, it goes, and sometimes it's occluded. The days of seeing the wrath of God in a prolonged drought or a heavy windstorm – believing that bad weather chastens our bad actions, in other words – are pretty much past. One sobering irony of global warming is the thought that it threatens to make weather moral again in a very different way. But these are thoughts too puzzling for the fine weather of these last few evenings, when it is almost impossible not to feel that this has come to us by right – as our due after a run of sticky days and as the best of what the month of June has to offer anyway. These are the nights for stoop sitting, not in long-suffering, as though we felt the curse of Cain on our shoulders, but like the young man and his dog I passed the other evening. Both sat quietly, watching the street. You could tell that what they were really doing was feeling the shape of the cool air around their bodies. It would have been a pleasure in itself, but it was all the more pleasurable for the memory of that hot spell. [...]

Figure 4: An excerpt of the news editorial "The Reward of Good Weather". This editorial neither affected the stance of annotators with conservative ideologies, nor the stance of those with liberal ideologies.

Exemplarily, Figure 3 shows an excerpt of an editorial on police brutality and misconduct from the corpus. This editorial challenged the stance of annotators with one ideology and reinforced the stance of those with the opposing ideology. Accordingly, it meets our conditions of being of high argumentation quality.

On the other hand, Figure 4 shows an excerpt of an editorial on global warming from the corpus. This editorial did not affect annotators of either ideology. As a result, it meets our conditions of being of low argumentation quality.

5 Analysis

This section provides insights into the annotations of our corpus. We first outline the reliability of the annotations in terms of inter-annotator agreement, and we compare the annotations of liberals and conservatives, highlighting the noteworthy differences. Then, we analyze the annotations regarding their argumentation quality according to our model.

5.1 Inter-Annotator Agreement

Table 4 lists agreement results for all annotators within each group (*liberals* and *conservatives*) as well as across both groups (*overall*).

The overall agreement is lower than the agreement within each group regarding the majority, full, and Krippendorff's α. For example, Krippendorff's α is 0.32 for liberals and 0.29 for conservatives, but only 0.16 overall. According to the Mann-Whitney test (Mann and Whitney, 1947), the difference be-

	Effect w/o Intensity			Effect vs. No Effect		
	Majority	**Full**	α	**Majority**	**Full**	α
Liberals	74%	33%	0.32	83%	48%	0.30
Conservatives	69%	20%	0.29	77%	31%	0.32
Overall	**64%**	**0%**	**0.16**	**72%**	**12%**	**0.17**

Table 4: Majority, full, and Krippendorff's α agreement for both political ideologies and overall for annotating *what* effect all news editorials in our corpus have (left side) and *whether* they have any effect (right side).

tween the effects annotated by liberals and those annotated by conservatives is significant at $p < 0.05$.

In general, the liberal group agreed more than the conservative, with a majority agreement of 74% against 69%, full agreement of 33% against 20%, and an α of 0.32 against 0.29. One reason for this may be given by the varying ideology distributions within each group: As mentioned before, the PEW political typology ranges from far right to far left, and in Figure 2, we see that the annotators with liberal ideology are further from the middle than the conservative annotators. An annotator closer to the middle is likely to be less devoted than an annotator with a more extreme ideology (e.g., *solid liberals* or *core conservatives*).

The observed α agreement can be interpreted as "fair" for both liberals and conservatives, and as "slight" overall. We see two reasons for this limited agreement: (1) The task at hand is very subjective, and (2) the distribution of labels is skewed. For instance, *challenging* is chosen significantly less than *no effect* and *reinforcing*. Krippendorff's α has been shown to be often low in such cases (Di Eugenio and Glass, 2004). Indeed, the values are in line with those obtained for similar tasks in other studies (Wachsmuth et al., 2017a).

5.2 Editorial Argumentation Quality

Table 5 shows the distribution of news editorials over their combined effect on the two opposing belief groups, ignoring which group is liberal and which is conservative. According to our model, 6% of the editorials are of high quality, 46% of medium quality, and 48% of low quality.

Only one of the 1000 editorials changed the stance of either group with majority. According to a one-sided binomial test, the proportion of readers changing their stance after reading an editorial is significantly lower than 1% at $p < 0.001$. This speaks for our hypothesis that editorials do not serve to persuade readers in most cases. By contrast, 151 editorials empowered annotators of either groups to argue better about the discussed issue, which shows the importance and applicability of the 'empower' notion in our model.

Quality	**Group A**	**Group B**	**#**	**%**
High	Challenging	Challenging	4	1%
	Challenging	Reinforcing	37	5%
Medium	Reinforcing	Reinforcing	338	44%
	Challenging	No Effect	19	2%
Low	Reinforcing	No Effect	296	38%
	No Effect	No Effect	76	10%
	Either group changed stance		1	0%
	Either group was empowered		151	20%

Table 5: Distribution of combined majority effects of the news editorials in our corpus on opposing belief groups, along with their quality according to our model. Editorials without majority agreement are ignored here.

296 editorials reinforced the stance of one group and did not affect the other group. From these, 244 reinforced the stance of liberal annotators, suggesting that their stance more often equals the stance of the editorials. This matches expectation, given that the New York Times is seen as a rather left news portal. According to a Fisher's exact test, the difference in choosing *reinforcing* between liberals and conservatives is significant at $p < 0.05$.

5.3 Personality Traits

Our model of argumentation quality is built by profiling readers based on their political ideologies, whereas Lukin et al. (2017) profiles the audience for the "Big Five" personality traits to see whether different personality types are more open to different types of arguments. Although we focus in this paper on the audience's political belief, we also expected useful insights from correlating the personality traits of our annotators to the editorials' effects. For this reason, we asked our annotators to take the personality test based on the "Big Five" (Goldberg, 1990).

Table 6 shows the counts of the personality traits of annotators based on their political orientations. Since our primary goal was to have annotators evenly distributed over their political orientations, we did not control the distribution based on personality traits.

We computed the correlations between the annotators' personality traits and their annotations regarding the effect of editorials. Table 7 shows Kendall's τ correlation coefficient between the

	"Big Five" Personality Traits														
	Agreeableness			**Conscientiousness**			**Extraversion**			**Neuroticism**			**Openness**		
	Low	Average	High	Low	Average	High	Low	Average	High	Low	Average	High	Low	Average	High
Liberals	3	2	7	2	6	4	4	3	5	8	2	2	3	2	7
Conservatives	5	0	7	4	3	5	8	2	2	2	3	7	4	6	2
Overall	**8**	**2**	**14**	**6**	**9**	**9**	**12**	**5**	**7**	**10**	**5**	**9**	**7**	**8**	**9**

Table 6: Counts of the annotators' "Big Five" personality trait values, depending on their political orientation.

Kendall's τ	**"Big Five" Personality Traits**				
	Agree.	**Consc.**	**Extra.**	**Neuro.**	**Openn.**
Liberals	0.02	0.04	*0.15	-0.06	0.06
Conservatives	0.14	-0.14	*0.23	0.02	*0.31
Overall	**0.10**	**-0.06**	**0.24**	**-0.11**	**0.22**

Table 7: Kendall's τ correlation between the annotators' "Big Five" personality traits and the effect of a news editorial depending on the annotators' political orientation. Values with * are discussed in Section 5.3.

"Big Five" (Goldberg, 1990) and the effect of an editorial on liberal and conservative annotators.

As discussed in Section 2, Lukin et al. (2017) found specific types of arguments to be persuasive for people with specific personality traits. Analogously, we found correlations between the effect of editorials and combinations of political ideology and personality traits of readers. For both liberals and conservatives, for instance, there is a positive correlation between their choices for editorial effect and the *extraversion* trait. This trait characterizes people who tend to be more dominant in social settings (Friedman et al., 2010). Similarly, for conservatives there is a positive correlation between their choices for editorial effect and the *openness* trait. This trait characterizes active imagination or high curiosity (Costa and McCrae, 1992).

5.4 Explanations

The annotators were asked to justify their answers, resulting in in a total of 6000 explanations. All explanations were manually inspected by us. Among others, we found that the majority of annotators selected *no effect* for two main reasons. One reason was that, as expected, an annotator already had a stance towards the discussed topic of the editorial, but he or she found the editorial rather ineffective. The second reason was that an annotator did not have a stance regarding the topic, but also expressed no interest in the topic.

6 Conclusion

This paper has presented a new model for the argumentation quality of news editorials. As its main dimensions the model combines the *reader's beliefs* and the *editorial effect*. While the reader's beliefs are defined in terms of the political orientation of a reader, the editorial effect is defined as an editorial's capability to either challenge readers with opposing stance or to empower readers with the same stance. This way, our model goes beyond approaches that aim at quantifying quality as an absolute value. Instead, we analyze how editorials increase or decrease the gap between readers with opposing beliefs (van Dijk, 1995).

To compute the determinants of our proposed model and to analyze its potential, we built a new corpus of 1000 editorials from the New York Times. Each editorial has been annotated regarding its perceived effect by three liberals and by three conservatives. Our analysis of the corpus provided first insights: As expected, readers with identical beliefs largely agree on the effect of editorials. In particular, we provide empirical evidence for the hypothesis that editorials rarely change the stance of readers. With our approach, we can quantify such effects more precisely.

We also observed that the ideology of the New York Times seems to be reflected in the annotated corpus: The editorials reinforced the stance of many annotators with liberal ideology, while they often had no effect on annotators with conservative ideology. In addition, we found correlations between the effects of editorials and the combination of political ideology and personality trait.

We consider the presented model and the new corpus as substantial resources for fostering research on computational argumentation. We ourselves plan to use these resources in future work, in particular, for developing computational approaches to assess argumentation quality.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

P. Anand, J. King, Jordan Boyd-Graber, E. Wagner, C. Martell, Douglas Oard, and Philip Resnik. 2011. Believe me — We can do this! Annotating persuasive acts in blog text. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Aristotle. translated 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.

Bal Krishna Bal. 2009. Towards an analysis of opinions in news editorials: How positive was the year? (project abstract). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 260–263. Association for Computational Linguistics.

Bal Krishna Bal and Patrick Saint Dizier. 2010. Towards building annotated resources for analyzing opinions and argumentation in news editorials. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212. Association for Computational Linguistics.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413. Association for Computational Linguistics.

Paul T Costa and Robert R McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Barbara Di Eugenio and Michael Glass. 2004. he Kappa Statistic: A Second Look. *Computational Linguistics*, 30(1):95–101.

Teun A. van Dijk. 1995. Opinions and Ideologies in Editorials. In *Proceedings of the 4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought*, Athens.

Carrol Doherty, Jocelyn Kiley, and Bridget Johnson. 2017. Political typology reveals deep fissures on the right and left doherty.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Howard S Friedman, Margaret L Kern, and Chandra A Reynolds. 2010. Personality and health, subjective well-being, and longevity. *Journal of Personality*, 78(1):179–216.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

H. Halmari and T. Virtanen. 2005. *Persuasion Across Genres: A linguistic approach*. Pragmatics & Beyond New Series. John Benjamins Publishing Company.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. International Debate Education Association.

Johannes Kiesel, Khalid Al Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 35–38. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics.

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using

argument-based features to predict and analyse review helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1363. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. What makes writing great? First experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753. Association for Computational Linguistics.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2017. Lightly-supervised modeling of argument persuasiveness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 594–604. Asian Federation of Natural Language Processing.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351. Asian Federation of Natural Language Processing.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence — An automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. Corpus number LDC2008T19. In *Linguistic Data Consortium, Philadelphia*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510. Dublin City University and Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International World Wide Web Conference*, pages 613–624.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136. Association for Computational Linguistics.