Corpus of News Articles Annotated with Article Level Subjectivity

Ahmet Aker, Hauke Gravenkamp, Sabrina J. Mayer
firstName.lastName@uni-due.de
University of Duisburg-Essen

Marius Hamacher, Anne Smets, Alicia Nti
firstName.lastName@stud.uni-due.de
University of Duisburg-Essen

Johannes Erdmann, Julia Serong, Anna Welpinghus
firstName.lastName@tu.dortmund.de
Technical University of Dortmund

Francesco Marchi
francesco.marchi@rub.de
Ruhr University of Bochum

ABSTRACT

Over the past few years, opinion mining or subjectivity scoring has been studied quite extensively, and technical solutions are proposed. However, so far, the subjectivity scoring is restricted to the level of short textual units such as sentences. A desired situation would be that there are also solutions computing subjectivity scores on the level of, e.g. the entire article. We applied a sentence-based subjectivity scorer on news articles and compared the results with the scores provided by human judges. Our comparison shows that there is a huge gap between machines and human judges when the task is to determine subjectivity scores at article level. To close this gap, we release a new human-annotated dataset containing 250 news articles with subjectivity scores annotated at article level. Each article is annotated by at least ten people. The articles are evenly divided into fake and non-fake categories. Our investigation on this corpus shows that fake articles are significantly more subjective than non-fake ones. The dataset will be made publicly available.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

corpus of news articles, subjectivity annotation, fake news

ACM Reference Format:
Ahmet Aker, Hauke Gravenkamp, Sabrina J. Mayer, Marius Hamacher, Anne Smets, Alicia Nti, Johannes Erdmann, Julia Serong, Anna Welpinghus, and Francesco Marchi. 2019. Corpus of News Articles Annotated with Article Level Subjectivity. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The web has never been as big as it is today. It contains a tremendous amount of information represented in the form of standard web documents, videos, images, blog posts, social media entries, and many others. One of the reasons for this massive growth is that the web is not anymore shaped by only a few experts or dedicated people and institutions but by everyone who has access to it. Although this new style of contribution towards the web content has led to an immense information richness, alternative views, and diversity, it has, however, also raised new challenges. It has stripped out the traditional information providers from their gate-keeping role [2] and has left the public in a jungle of web content with varying quality, reliability, veracity, and credibility. A web user wandering around in this jungle is likely to be misled, manipulated in her belief towards a specific group's interest, political party, a theory, etc. and psychologically attacked to capture the consumer's attention and lead her towards actions not only harmful to herself but also to society. A famous example involving harmful action is the "Pizzagate" scandal, which was provoked by misinformation shared on social media about 2016 presidential candidate Hillary Clinton's alleged connection to a child pornography ring acting in a pizzeria that ended up with gun shootings [1, 3].

Fuhr and colleagues [5] discuss the idea of implementing information nutrition labels for news articles as a means to fight against online misinformation such as fake news. They propose to label every online news article with information nutrition labels that describe the ingredients of the article and give readers a chance to make an informed judgment about what they are reading. The authors discuss nine different labels, including the assignment of subjectivity scores to the articles. The aim of a subjectivity score is to capture the level of to what extent the statements in the article have been influenced by the author's personal viewpoint. A subjective statement has a basis in reality but reflects the perspective through which the writer views the topic. Such statements are difficult to verify using facts and figures. Fuhr et al. [5] assume that misinformation such as fake articles tend to be more subjective than non-fake ones.

Related work has proposed methods to determine subjectivity scores automatically [8, 10]. These approaches compute subjectivity scores on the sentence level. Using this paradigm, one can compute an article level subjectivity score by aggregating the sentences' subjectivity scores and averaging these values [7]. However, our analysis shows that such an approach does not correlate with article level subjectivity scores provided by humans. Human readers tend to rate an article as highly subjective even if only a few (sometimes even just one or two) are colored with authors' viewpoints, and the remaining sentences are non-subjective. To close this gap, we

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

argue that automated solutions should also work directly on the entire article's content.

To enable such a paradigm, we release a dataset containing 250 news articles with article level subjectivity labels. This new corpus should help systems learn to compute subjectivity scores at article level. To our knowledge, this is the first corpus with article level subjectivity scores. Our articles are split into fake (125) and non-fake (125) articles. We show that at the article level, fake articles are significantly more subjective than non-fake ones. This finding supports the assumption of Fuhr et al. [5] and we believe that subjectivity will help readers to distinguish between credible and non-credible articles when performed at article level.

In the following, we will first describe the dataset annotated with subjectivity at article level (Section 2). In Section 3, we present inter-rater agreement among the annotators, the analysis of subjectivity provided for fake and non-fake articles, as well as a qualitative analysis of articles with low and high subjectivity scores. In Section 4, we give results on our correlation analysis between human subjectivity scores and those obtained automatically. Finally, we discuss our findings and conclude the paper in Section 5.

2 DATASET

We retrieved the news articles annotated in this work from *FakeNewsNet* [11], a corpus of news stories divided into fake and non-fake articles. To determine whether a story is fake or not, the *FakeNewsNet* authors extracted scores and articles from two prevalent fact checking sites *PolitiFact*[1] and *GossipCop*[2]. We sampled 125 fake and non-fake articles from this corpus. All articles are about political news, mostly the 2016 US presidential elections. Table 1 lists textual statistics about the articles.

Table 1: Textual statistics about articles in the dataset.

		fake	non-fake
text length	*min*	820	720
	max	10062	12959
	median	2576	3003
	mean	2832.4	4124.3
sentences	*min*	6	6
	max	88	144
	median	22	27
	mean	24.5	36.1
average words per sentence	*min*	11.048	8.035
	max	35.7	36.7
	median	19.8	19.5
	mean	20.7	19.9

Each news article was rated between 10 and 22 times ($mean = 15.524, median = 15$) and each annotator rated 1 to 250 articles ($mean = 42.185, median = 17$). Annotators were recruited from colleagues and friends and were encouraged to refer the annotation project to their acquaintances. They were free to rate as many articles as they liked and were compensated with 3.5€ (or 3£ if they were residents of the UK) per article.

[1]https://www.politifact.com/
[2]https://www.gossipcop.com/

For annotating the articles, we used a custom online platform. After registration and reading the consent form, annotators were presented with annotation instructions and could start annotating. Articles were presented in a plain, text-only format and were accessible while answering the subjectivity questions.

Subjectivity was rated in two different ways. First, annotators were asked to rate textual qualities of the given article that indicate subjectivity; for instance, *The authors express their individual thoughts, beliefs, or attitudes.* These qualities were measured by five properties on a *5-Point Rating Scale*, labelled *Strongly Disagree* to *Strongly Agree*. Afterwards, annotators were asked to rate the subjectivity directly on a percentage scale (*Overall, how subjective is this article? Judge on a scale from 0-100.*). The value 0 indicates complete objectivity, 100 complete subjectivity.

This two-fold annotation approach was chosen to generate subjectivity scores that could be used to train models as well as subjectivity indicators that could provide insights as to why and how people rate the subjectivity of an article. In the present work, however, we only analyze the percentage scores for subjectivity. When discussing annotations, we refer to these subjectivity scores.

3 ANALYSIS OF SUBJECTIVITY SCORES

We first measure differences in inter-rater agreement for both article types to assess whether the annotators agree on the judgments or not. We also analyze the distribution of subjectivity ratings for fake and non-fake articles. Afterwards, we look at articles with particularly high or low scores to find differences in writing that could influence annotators in their ratings and determine whether an article is perceived as subjective.

3.1 Inter-rater Reliability Analysis

Inter-rater reliability is measured using the *Intra Class Correlation (ICC) Index*. It measures how closely values in different groups resemble each other. *ICC* scores range between 0 and 1. A score closer to 1 indicates that values within groups tend to be close to each other, whereas a score of 0 indicates that values take on vastly different values. In this case, groups are articles and values are annotation scores. A one-way random effects model for absolute agreement with average measures as observation units is assumed (*ICC(1,k)*). (We followed the guidelines of [6, 9] to select the ICC model parameters.)

Since not every annotator annotated every article, they are assumed to be a random effect in the model. We chose the minimum number of available annotations per article ($k = 10$) as the basis for the reliability analysis. In cases where more than 10 annotations were available for an article, we chose 10 annotations randomly. Observational units are average measures since the subjectivity for each article represents the average of all human annotations for the given article.

The total ICC is .89, which indicates good to excellent reliability [9]. Reliability is slightly higher for real articles ($ICC(1, 10) = .88$) than for fake ones ($ICC(1, 10) = .77$) (see Table 2). Note that there is a large discrepancy between the average point estimates and the single point estimates ($ICC(1, 1) = .46, CI[.95] = [.41, .51]$) for the same data. While this is generally expected [6], we considered the difference to be large enough to report.

Table 2: Intraclass Correlation Values

					95% CI	
	N	Raters	Unit	ICC	Lower	Upper
total	250	10	average	.89	.87	.91
			single	.46	.41	.51
fake	125	10	average	.77	.71	.83
non-fake	125	10	average	.88	.85	.91

3.2 Annotation Distribution

The dataset contains 3881 subjectivity score annotations, ranging between 0 and 100. The mean score is 54.86 with a standard deviation of 34.31. Including all articles, scores are mostly uniformly distributed with minor peaks at the max and min values (Figure 1).

The distribution changes when fake and non-fake articles are analyzed separately. Fake articles receive higher scores ($mean = 68.38$) than non-fake ones ($mean = 41.30$). We found a significant difference ($t(3879) = 26.74, p < .001$) of large magnitude ($cohen's\ d = 0.86$) between the two, using a *t-test*. Also, the percentage of fake articles with a subjectivity score of 50 or more stands at 77.3 compared to real articles where only 43.8 percent received a score over 50. This analysis shows that indeed fake articles are rated significantly more subjective than the non-fake ones.

3.3 Qualitative Analysis

We analyzed fake and real articles with the highest and lowest subjectivity scores qualitatively to find clues as to why raters assigned extraordinarily high or low scores.

Our preliminary analysis indicates differences in language use. However, differences between fake and real articles appear mostly in articles with high subjectivity scores. Real articles contain more figure and fact listing, in part for comparisons, when compared with fake articles. For instance, *"ABC drew 13.5 million viewers, CBS drew 12.1 million, Fox News drew 11.4 million, CNN drew 9.9 million [...]"* or *"Trump garners just 7% support compared to Clinton's 81% with the voting bloc."* However, this distinction might be a coincidence. Nevertheless, both fake and real articles make use of quotations and third person narrative style in indirect speech. To state an example: *"'We have to be optimistic that we'll see some real concrete commitments' said Amnesty International's Interim Executive Director Margaret Huang"* or *"Obama said he [...]"*.

All in all, low subjectivity rated articles seem to be re-narrating and therefore have less subjective bias and opinions included.

In general if articles are highly subjective, regardless whether they come from fake or non-fake category they seem to make use of colloquial writing and stylistic devices more frequently.

Moreover, the first person narrative and the use of pronouns is recognizable in the articles. For example: *"I envision [...]"*, *"I hope [...]"*, or *"[...] I can tell you from the facts that the story is a little bit different[...]"*. Furthermore, statements that give distinct opinions and judgment such as *"[...] I can't believe this nonsense"*, *"Are you kidding me? That's ridiculous"*, and *"Trump's cheap explanation sounded like something a third grader on the playground would offer rather than a presidential candidate"* are found within some of the articles. The latter quotation also provides an example for comparisons

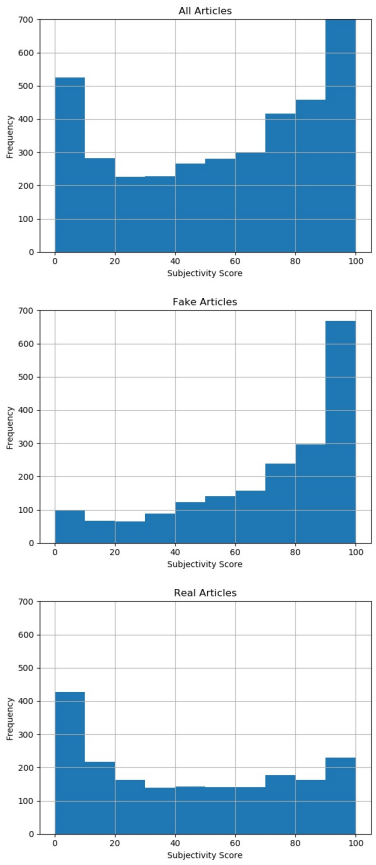

Figure 1: Subjectivity scores.

and analogies that are present in some text passages. Moreover, we found that polemic and aggressive language is used in rated articles: *"Donald Trump Jr. Is quite possibly the dumbest person ever to take to social media."* Furthermore, some parts include direct appellation like *"You are going to be mad that I told you this. You're going to wonder where the toughness of America went and what happened to make people think being a p*ssy is the way to be"* or *"But hey, who needs freedom, right?"*. The latter example is also written in a sarcastic way, which is also identifiable in other text parts, e.g. *"Meanwhile, President Trump tweeted his thoughts and prayers to McCain because they are Christians who love everyone even if they have had their differences in the past."* The use of imperatives, as like *"IMPEACH DONALD TRUMP!"* and *"Forget him"* is another feature found in the articles, which in fact gives reference to subjectivity exclusively by its sentence. Apart from this, some high-subjectivity articles include rhetorical devices like analogies, comparisons, and rhetorical questions, which do not occur in the articles with the lowest scores in the same manner.

4 COMPARISON: PREDICTIONS VS. HUMANS

To see how existing sentence-level subjectivity analysis models perform on the dataset, we use the *Pattern*[3] Web Mining Package [4].

[3]https://github.com/pattern3

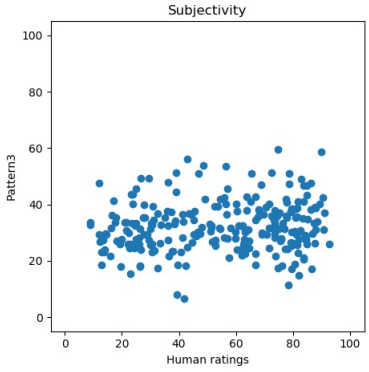

Figure 2: Human ratings and model predictions.

The package provides a dictionary-based subjectivity analyzer with a dictionary of adjectives and their corresponding sentiment polarity and subjectivity scores. Model predictions were obtained by processing the articles in the dataset sentence by sentence and averaging over the sentence scores. Human ratings represent the average subjectivity score per article.

Figure 2 plots human ratings and the model predictions. The correlation is significant yet small ($r = .126, R^2 = .016, p = .046$), prediction errors are high ($MSE = 1104.35, MAE = 27.59$).

Generally, model predictions are lower than the human ratings. The model predictions range from 6.7 to 59.49 with a mean of 32.14 and a standard deviation of 8.85, whereas the human annotations span a wide range of values, rating articles' subjectivity from 8.75 to 92.65, with a mean of 54.67 and a standard deviation of 23.97.

We also looked at the distribution of subjectivity scores (as in Section 3.2) for the model predictions. When comparing scores assigned to fake articles ($mean = 33.54$) and scores assigned to real articles ($mean = 30.73$), the predictions do not differ significantly ($t(248) = 2.54, p = .065, cohen's\ d = 0.32$). This indicates that computing subjectivity scores as in this setting (performing sentence level subjectivity score computation) is not useful for fake news detection. On the other hand, the human ratings on article level, where we found a significant difference between fake ($mean = 68.10$) and real ($mean = 41.24$) articles with a large magnitude ($t(248) = 10.69, p < .001, cohen's\ d = 1.35$), are indeed a useful feature for distinguishing between fake and non-fake articles.

5 DISCUSSION AND CONCLUSION

A novel human annotated subjectivity dataset is presented in this paper.[4] To the best of our knowledge, it is the first dataset providing high quality article level subjectivity ratings.

Our analysis of model predictions shows that sentence level subjectivity estimates are unable to match human estimates for entire articles. Sentence level models underestimate true subjectivity scores, probably due to the fact that results are averaged over all sentences. If an article contains mostly objective sentences and only a few sentences with highly subjective statements, these models will assign a relatively low score to the article. Contrarily, for human readers, a few of such opinionated statements can shape the perception of the entire article. Subjectivity models should therefore be trained and operate on the article level rather than on the sentence level (if the unit of analysis is on the article level). Our dataset can be used to train such models and thus is a valuable addition to the collection of available subjectivity datasets.

Furthermore, fake and real articles differ in the distribution of subjectivity annotations. Real articles in our dataset receive significantly lower subjectivity scores than fake ones. This finding qualifies subjectivity as a potential feature for fake news classification of political news articles. Sentence level models failed to generate scores that reflect this relation. Models could be improved by making predictions on the article level and by using our dataset for training.

Future research could be aimed at examining this finding further by incorporating more articles, potentially also from different topic domains, as our dataset includes only political news articles.

We started investigating, where differences in subjectivity may be coming from and (unsurprisingly) find that more extreme and emotionally charged statements were used in highly subjective articles. As mentioned earlier, the interesting finding here is that even a few such statements seem to affect the overall impression of an article's subjectivity.

In future studies, this investigation could be expanded by quantitative analyses, aiming at finding an impact rating of sentences in articles or qualitatively aiming at finding factors that influence people's perception of subjectivity in an article.

6 ACKNOWLEDGEMENTS

This work was funded by the Global Young Faculty[5] and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media"

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. *arXiv preprint arXiv:1810.01765* (2018).
[3] Hal Berghel. 2017. Lies, damn lies, and fake news. *Computer* 2 (2017), 80–85.
[4] Tom De Smedt and Walter Daelemans. 2012. Pattern for Python. *J. Mach. Learn. Res.* 13, 1 (June 2012), 2063–2067. http://dl.acm.org/citation.cfm?id=2503308.2343710
[5] Norbert Fuhr, Wolfgang Nejdl, Isabella Peters, Benno Stein, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, and Josiane Mothe. 2018. An Information Nutritional Label for Online Documents. *ACM SIGIR Forum* 51, 3 (feb 2018), 46–66. https://doi.org/10.1145/3190580.3190588
[6] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8, 22833776 (2012), 23–34. https://www.ncbi.nlm.nih.gov/pmc/PMC3402032/
[7] Vincentius Kevin, Birte Högden, Claudia Schwenger, Ali Sahan, Neelu Madan, Piush Aggarwal, Anusha Bangaru, Farid Muradov, and Ahmet Aker. 2018. Information Nutrition Labels: A Plugin for Online News Evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 28–33. https://www.aclweb.org/anthology/W18-5505
[8] Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts.*
[9] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic*

[4] https://github.com/ahmetaker/newsArticlesWithSubjectivityScores

[5] https://www.global-young-faculty.de/

medicine 15, 27330520 (June 2016), 155–163. https://www.ncbi.nlm.nih.gov/pmc/PMC4913118/

[10] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 415–463.

[11] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *CoRR* abs/1809.01286 (2018). arXiv:1809.01286 http://arxiv.org/abs/1809.01286