Corpus of news articles annotated with article-level sentiment

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

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

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

Francesco Marchi
Ruhr University Bochum, Germany
firstName.lastName@rub.de

Abstract

Research on sentiment analysis is in its mature status. Studies on this topic have proposed various solutions and datasets to guide machine-learning approaches. However, so far the sentiment scoring is restricted to the level of short textual units such as sentences. Our comparison shows that there is a huge gap between machines and human judges when the task is to determine sentiment scores of a longer text such as a news article. To close this gap, we propose a new human-annotated dataset containing 250 news articles with sentiment labels at article level. Each article is annotated by at least 10 people. The articles are evenly divided into fake and non-fake categories. Our investigation on this corpus shows that fake articles are significantly more sentimental than non-fake ones. The dataset will be made publicly available.

In: A. Aker, D. Albakour, A. Barrón-Cedeño, S. Dori-Hacohen, M. Martinez, J. Stray, S. Tippmann (eds.): Proceedings of the NewsIR'19 Workshop at SIGIR, Paris, France, 25-July-2019, published at http://ceur-ws.org

1 Introduction

Nowadays, the amount of online news content is immense and its sources are very diverse. For the readers and other consumers of online news who value balanced, diverse, and reliable information, it is necessary to have access to additional information to evaluate the available news articles. For this purpose, Fuhr et al. [7] propose to label every online news article with information nutrition labels to describe the ingredients of the article and thus give the reader a chance to evaluate what she is reading. This concept is analogous to food packages where nutrition labels help buyers in their decision-making. The authors discuss 9 different information nutrition labels including sentiment. The sentiment of a news article is subtly reflected by the tone and effective content of a writer's words [5]. Fuhr et al. [7] conclude that knowing about an article's level of sentiment could help the reader to judge the credibility and whether it is trying to deceive the reader by relying on emotional communication.

Sentiment analysis is a mature research direction and has been summarized by several overview papers and books [13, 3, 4]. Commonly, sentiment is computed on a small fraction of text such as a phrase or sentence. Using this strategy, authors of [11, 14, 1] analyze for instance Twitter posts. To compute sentiment over a text, such as a news article that spans over several sentences, [9] use the aggregated average sentiment score of the text's sentences. However, our current study shows that this does not align with the human perception of sentiment. If there are only, e.g. two sentences in the article which are sentimentally loaded and the remaining sentences are neutral, a sentence-based sentiment scorer will label the article as not sentimental or will assign a low sentiment score. On the contrary, our study shows that humans may consider the entire article as highly sentimental even if there are only 1-2 sentences that are highly sentimental.

In this work, we propose to release a dataset containing 250 news articles with article-level sentiment labels.[1] These labels were assigned to each article by at least 10 paid annotators. To our knowledge, this is the first article-level sentiment labeled corpus. We believe this corpus will open new ways of addressing the sentiment perception gap between humans and machines. Over this corpus, we also run two automatic sentiment assessors and show that their scores do not correlate with human-assigned scores.

In addition, our articles are split into fake (125) and non-fake (125) articles. We show that at the article level, fake articles are significantly more sentimental than the non-fake ones. This finding supports the assumption that sentiment will help readers to distinguish between credible and non-credible articles.

In the following, we will first describe the dataset annotated with sentiment at article level (Section 2). In Section 3, we present inter-rater agreement among the annotators, the analysis of sentiment provided for fake and non-fake articles, as well as a qualitative analysis of articles with low and high sentiment scores. In Section 4, we provide results about our correlation analysis between human sentiment 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* [15], 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 articles and veracity scores from two prevalent fact-checking sites *PolitiFact*[2] and *GossipCop*[3]. We sampled 125 fake and 125 non-fake articles from this corpus. All articles are dealing with political news, mostly the 2016 US presidential election. Table 1 lists textual statistics about the articles.

Each news article was rated between 10 and 22 times ($mean = 15.524, median = 15$) and each annotator rated 1 to 250 articles ($mean = 42.185, median = 17$).

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	4124.4
sentences	*min*	6	6
	max	88	144
	median	22	27
	mean	24.4	36.1
sentence average words	*min*	11.0	8.0
	max	35.7	36.7
	median	19.8	19.5
	mean	20.6	19.9

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. The recruitment method and relatively high monetary compensation were chosen to ensure high data quality.

Sentiment was rated in two different ways. First, annotators were asked to rate textual qualities of the given article that indicate sentiment, for instance, *The article contains many words that transport particularly strong emotions.*. These qualities were measured by five properties on a *5-Point Rating Scale*, labeled *Strongly Disagree* to *Strongly Agree*. Afterwards, annotators were asked to rate sentiment directly on a percentage scale (*Overall, how emotionally charged is the article? Judge on a scale from 0-100*), 100 indicating high sentiment intensity and 0 indicating low sentiment intensity.

We opted for the two-fold annotation approach to generate sentiment scores that could be used to train machine-learning models as well as sentiment indicators that could provide insights as to why and how people rate the level of sentiment of an article. In the present work, however, we only analyze the percentage scores for sentiment. When referring to annotations, we refer to these sentiment scores. The other sentiment variables are not discussed in this current work due to spacial constraints.

Note that annotators did not annotate the sentiment *polarity*, e.g. *"highly positive"* or *"slightly negative"*, but only the sentiment *intensity*, e.g. *"high"* or *"low"*. In this scheme, highly positive and highly negative articles receive the same score. We chose this annotation scheme since article-level polarity seems

[1] https://github.com/ahmetaker/newsArticlesWithSentimentScore

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

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

less informative for an entire article: In cases where a single article praises one position and condemns another, giving an overall polarity score is ambiguous and sentence-level polarity scores may be more informative.

The notion of sentiment intensity is still different from *subjectivity.* A subjective statement contains personal views of an author whereas an objective article contains facts about a certain topic. Both subjective and objective statements may or may not contain sentiment [12]. For example, *"the man was killed"* expresses a negative sentiment in an objective fashion, while *"I believe the earth is flat"* is a subjective statement expressing no sentiment. For an investigation of article level subjectivity, see [2].

3 Analysis of Sentiment Scores

First, we measure differences in inter-rater agreement for fake and non-fake articles in order to see whether the annotators agree on the judgments or not. We also analyze the distribution of sentiment ratings to see whether there are differences in sentiment scores for fake and non-fake articles. Afterwards, we look at articles with particularly high or low sentiment scores to find differences in the writing of the articles that could influence annotators in their ratings and determine whether an article is perceived sentimental.

3.1 Inter-rater Reliability Analysis

Inter-rater reliability is measured using the *Intra Class Correlation (ICC) Index.* 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 [8, 10] to select the ICC model parameters.)

Since not every annotator annotated every article, annotators 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 randomly chose 10 annotations. Observational units are average measures since the sentiment for each article is going to be the average of all human annotations for the given article.

The total Intra Class Correlation is 0.88, which indicates good to excellent reliability [10]. Reliability is slightly higher for real ($ICC(1, 10) = .90$) than for fake articles ($ICC(1, 10) = .76$) (see Table 2).

Note that there is a large discrepancy between the average point estimates and the single point estimates for the same data ($ICC(1, 1) = .42, CI[.95] = [.37, .48]$). While this is generally expected [8], 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	.88	.86	.90
			single	.42	.37	.48
fake	125	10	average	.76	.67	.81
non-fake	125	10	average	.90	.87	.92

3.2 Annotation Distribution

The dataset contains 3788 sentiment score annotations, ranging between 0 and 100. The mean score is 49.92 with a standard deviation of 32.54. When looking at all articles, scores are mostly uniformly distributed with minor peaks at the maximum and minimum values (see Figure 1). The distribution changes when dividing the articles into fake and real ones. Fake articles receive higher scores ($mean = 61.50$) than non-fake ones ($mean = 38.69$). We found a significant difference ($t(3786) = 22.99, p < .001$) of medium magnitude ($cohen's\ d = .75$), using a *t-Test.* In addition, the percentage of fake articles with a sentiment score of 50 or higher stands at 70.4 compared to real articles where only 40.6 percent were rated with a score above 50. This shows that indeed fake articles are rated significantly more sentimental than the non-fake ones.

3.3 Qualitative Analysis

A first qualitative analysis of the articles rated with the highest and lowest mean sentiment scores indicates differences in language use and sentence structure.

Articles with a low sentiment score are mostly election reports and contain listings of facts and figures. To give examples: *"Solid Republican: Alabama (9), Alaska (3), Arkansas (6), Idaho (4), Indiana (11), Kansas (6), Kentucky (8), Louisiana (8) [...]"*, or *"Clinton's strength comes from the Atlanta area, where she leads Trump 55% to 35%. But Trump leads her 51% to 33% elsewhere in the Peach state. She leads 88% to 4% among (..)."*

The last example also demonstrates the use of a repetitive and simple sentence structure, for instance, the iterating use of the word *leads. "In Iowa Sept. 29. In Kansas Oct. 19. [...]"* states another example for the repeated use of language. On the whole, the used language seems unemotional, rather neutral and without bias.

Articles with the highest mean score seem to consist of a larger number of negative words. *"Kill", "murder", "guns", "shooting", "racism"* and *"dead and bloodied"* are a few specific examples of negative words we observed in the articles. To some extent, offensive language is used which indicates a subjective view and

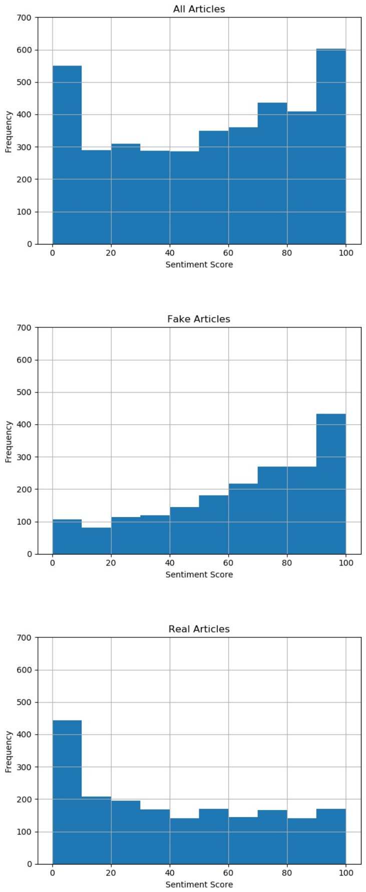

Figure 1: Histograms of sentiment scores. Values are sorted into 10 categories

bias. Statements such as *”[...] sick human being unfit for any political office [...]”*, or *”[...] nothing but a bunch of idiot lowlifes”* can be quoted as exemplary for offensive language use.

In some high-sentiment articles, we also found rhetorical devices such as analogies, comparisons, and rhetorical questions, which do not occur in the same manner in the low-sentiment articles. Analogies and comparisons are initiated by the word *like* such as in the following sentence: *”Clinton speculated about this, and like a predictable rube under the hot lights Trump cracked under the pressure.”* The following sentence gives an example for a rhetorical question found in one of the articles:*”Did Trump say he was interested in paying higher taxes? No. Did Trump say he would like to reform the tax code so that he would be forced to pay higher taxes? No.”*

4 Comparison between Model Predictions and Human Annotations

To see how existing sentence-level sentiment analysis models perform on the dataset, we used the *Pattern3*[4] Web Mining Package [6] and the *Stanford Core NLP*[5] Package.

The *Pattern3* package provides a dictionary-based sentiment analyzer with a dictionary of adjectives and their corresponding sentiment polarity and intensity. The model determines the sentiment score of a sentence by averaging the sentiment scores of all adjectives in a sentence. Scores range between -1.0 (negative) and 1.0 (positive).

The *Stanford Core NLP* package provides a recursive neural network model for sentiment analysis [**?**]. It assigns sentiment labels based on the contents and syntactic structure of a sentence. The output is one of five labels (*very negative, negative, neutral, positive, very positive*).

Model predictions were obtained by processing the articles in the dataset sentence by sentence and averaging over the sentence scores. Since the models assign sentiment values on different scales than the one used by our annotators, we mapped the values to match our scale. For the *Pattern3* scores, we took the absolute value and multiplied it by 100 and for *Stanford* scores, we mapped the labels to intensity scores (*very negative* $= 100$, *negative* $= 50$, *neutral* $= 0$, *positive* $= 50$, *very positive* $= 100$).

Human ratings represent the average sentiment score per article.

[4] https://github.com/pattern3
[5] https://stanfordnlp.github.io/CoreNLP/

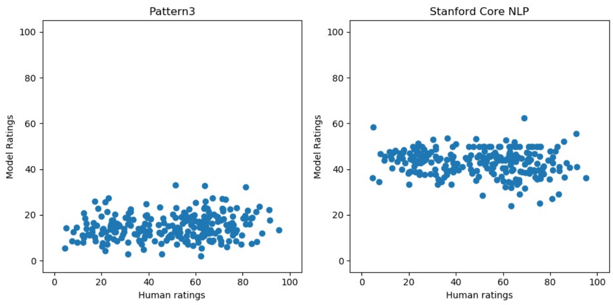

Figure 2: Scatter plot showing human ratings and model predictions for each sentiment analyzer.

4.1 Results

In general, model predictions are lower than the human ratings and span a more narrower of values. Model predictions of the *Pattern3* Sentiment Analyzer range from 2.04 to 32.99 with a mean of 14.81 and a standard deviation of 5.47. Predictions of the *Stanford Core NLP* Analyzer range from 24.07 to 62.5 with a mean of 43.18 and a standard deviation of 5.69. On the other hand, human annotations span a wide range of values from 4.55 to 95.25, with a mean of 49.39 and a standard deviation of 21.77.

Figure 2 shows a scatter plot of the human ratings and the model predictions. The correlations are significant yet very small ($r = .171, R^2 = .029, p < .001$ for *Pattern3*, $r = -.139, R^2 = .019, p = .028$ for *Stanford Core NLP*) and prediction errors are high, while those of *Pattern3* are larger ($MSE = 1657.87, MAE = 35.05$) than those of *Stanford Core NLP* ($MSE = 576, MAE = 20.42$). We also looked at the distribution of sentiment scores for the model predictions. When comparing scores assigned to fake articles ($mean = 15.19$) and scores assigned to real articles ($mean = 14.43$), the predictions do not differ significantly ($t(248) = 1.09, p = .28, cohen's\ d = 0.14$). On the other hand, analyzing the scores assigned by human annotators on the article level, we found a significant difference between fake articles ($mean = 60.36$) and real articles ($mean = 38.42$) with a large magnitude ($t(248) = 9.21, p < .001, cohen's\ d = 1.17$). The results indicate that the computation of an overall sentiment score based on sentence-level sentiment scores is not useful for fake news detection. However, human ratings at article level can indeed be used to distinguish between fake and non-fake articles.

5 Discussion and Conclusion

A new human annotated sentiment dataset is presented in this paper. To the best of our knowledge, it is the first dataset providing high quality, article-level sentiment ratings.

Our analysis of model predictions shows that sentence-level sentiment estimates are unable to match human estimates for entire articles. Sentence-level models underestimate true sentiment scores, probably due to the fact that results are averaged over the sentiments of all sentences. The fact that the *Pattern3* predictions are generally lower than the ones of *Stanford Core NLP* supports this hypothesis, as *Pattern3* averages over all adjectives in a sentence and all sentences, whereas the *Stanford* model is only averaged over all sentences in the article. If an article contains mostly neutral sentences and only a few sentences with strong emotional statements, these models will assign the article a relatively low score. Contrarily, for human readers, already a few of such emotionally-charged sentences can shape the perception of the entire article. Sentiment analysis models should, therefore, operate at the article level rather than at the sentence level. Our dataset can be used to train such models and is thus a valuable addition to the collection of available sentiment datasets.

Furthermore, fake and real articles differ in the distribution of sentiment annotations. Real articles in our dataset receive significantly lower sentiment scores than fake ones. This qualifies sentiment 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 sentiment may be coming from and (unsurprisingly) find that more extreme and emotionally-charged statements were used in high-sentiment 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 sentiment.

In future studies, this investigation could be expanded either by detecting which sentences have the largest impact on the overall sentiment score of an article or by identifying individual-level determinants that affect people's perception of sentiment in an article.

6 ACKNOWLEDGEMENTS

This work was funded by the Global Young Faculty[6] and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media".

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

References

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (2011), pp. 30–38.

[2] Aker, A., Gravenkamp, H., Mayer, Sabrina, J., Hamacher, M., Smets, A., Nti, A., Erdmann, J., Serong, Julia Welpinghus, A., and Marchi, F. Corpus of news articles annotated with article level subjectivity. In *ROME 2019: Workshop on Reducing Online Misinformation Exposure* (2019).

[3] Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. *A practical guide to sentiment analysis*. Springer, 2017.

[4] Cambria, E., Schuller, B., Xia, Y., and Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems 28*, 2 (2013), 15–21.

[5] Conroy, N. J., Rubin, V. L., and Chen, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology 52*, 1 (2015), 1–4.

[6] De Smedt, T., and Daelemans, W. Pattern for python. *J. Mach. Learn. Res. 13*, 1 (June 2012), 2063–2067.

[7] Fuhr, N., Nejdl, W., Peters, I., Stein, B., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., and Mothe, J. An information nutritional label for online documents. *ACM SIGIR Forum 51*, 3 (feb 2018), 46–66.

[8] Hallgren, K. A. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology 8*, 22833776 (2012), 23–34.

[9] Kevin, V., Högden, B., Schwenger, C., Sahan, A., Madan, N., Aggarwal, P., Bangaru, A., Muradov, F., and Aker, A. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (Brussels, Belgium, Nov. 2018), Association for Computational Linguistics, pp. 28–33.

[10] Koo, T. K., and Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine 15*, 27330520 (June 2016), 155–163.

[11] Kouloumpis, E., Wilson, T., and Moore, J. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media* (2011).

[12] Liu, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*, 1 (2012), 1–167.

[13] Mejova, Y. Sentiment analysis: An overview. *University of Iowa, Computer Science Department* (2009).

[14] Pak, A., and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (2010), vol. 10, pp. 1320–1326.

[15] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR abs/1809.01286* (2018).