9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

How Faithful are Social Media Posts in Representing News Articles?

Keywords: Social Media, Fact-Checking, Rumour Detection, Twitter, COVID-19

Introduction — Social media has been gaining popularity as a source of news.[1] While these platforms serve as a convenient place to gain topical knowledge, many posts contain information that is either fake or misleading, and the COVID-19 pandemic has made the problem of misinformation proliferation a public health issue[2] as well.

This has lead to the development of automatic fact-checking, where the objective is to classify unsubstantiated claims as being true or false, typically by gathering evidence from a web search or a known source of reliable information, and checking the veracity of the claims against this evidence.[1] However, social media posts do not necessarily contain factual claims, and even when they reference a specific piece of news as evidence, the post can still lead to a spreading of misleading information through unfaithful representations of the evidence.[2]

In this study, we comprehensively analyze Twitter posts that link to COVID-19 related news articles, and create a dataset with tweets annotated with how the news is being represented faithfully. Finally, we propose a method of predicting faithfulness using a BERT-based model.

Analysis of Tweets with News Article Mentions — We crawled Japanese Twitter data to extract all tweets that contained a link to an external news website (source tweets), and then ranked the tweets by number of reactions. We then analyzed the obtained tweets.

During tweet extraction, we first filter unrelated tweets, then obtain the number of reactions associated with each source tweet by constructing a tree of reaction tweets, where the root node is the source tweet, and each leaf is a reply, retweet, or quote retweet that references the source tweet or any of the other leaves. An example of such a reaction tree is depicted in Figure 1. The number of total reactions was obtained by counting the number of nodes in the tree, after which we ranked source tweets and tweet groups by number of reactions.

To verify our method, we obtain 23,508 source tweets from May 1, 2022, with reaction counts shown in Figure 2. This indicates that sampling a limited number of source tweets (top 5 source tweets from the top 200 articles, comprising 2.55% of source tweets and 93.52% of all reactions on May 1) sufficiently covers influential tweets during that time period.

Annotation Schema — We construct an annotation scheme that captures the relationship between the source tweet and the referenced news article. Particular consideration is given to expressing the faithfulness with which the content of the news article has been represented in the tweet, by a combination of a category label expressing the tweet content, and an alteration label expressing the degree of alteration for that category.

Category labels are i) summarization/paraphrasing, ii) external reference, iv) opinion and v) title repetition. Tweets labeled summarization/paraphrasing are summarizing or paraphrasing a statement of fact that is in the article, Tweets labeled external reference contain a statement of fact that is likely supported by an external source. Tweets labeled opinion contain the opinions of the user. Tweets with title repetition contain a repetition of the article title.

For each relevant category, the following alteration labels are also given: i) no alteration, ii) minor alterations, ii) major alterations, and iv) fabrication. Tweets labeled no alteration represent the article content exactly. Tweets labeled minor alterations represent the article in a unique

[1] `https://pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/`
[2] `https://www.who.int/health-topics/infodemic`

way, although unlikely to produce misleading interpretations. Tweets labeled major alterations contain an alteration that misrepresents the article content, or misleads readers. Tweets labeled fabrication are completely unfaithful.

During annotation, the annotator must identify and label both the category and alteration type for each tweet. If applicable, multiple categories may be attributed to each tweet.

Dataset Overview — We leveraged COVID-19 related tweets collected during October 1-4, 2022 to obtain a total of 914 tweets which were subsequently annotated. Table 1 shows the label distribution of categories. According to this table, the category label with the most tweets was "Title Repetition," which can be explained by website features that allow users to 'share' articles on Twitter by pre-populating an empty post with the article title and link. This was followed by "Opinion" and "Summarization/paraphrasing," reflecting the frequency by which users will summarize the content of the article and/or express their own opinions in the tweet.

Focusing on tweets with the "Sum./Paraphrasing" label (Table 2), we see that approx. 10% of tweets have a degree of alteration that is at least misleading, which demonstrates the importance of considering the faithfulness of representations of news articles on social media, even when there is a direct link to the source. We also note that the 13.3% of tweets with "External Reference" is a possible target for fact-checking research.

To use the dataset in our task, we reorganize each label into "Subjectivity" and "Information Inheritance" label classes. Details on the labels and the criteria with which each label is applied are outlined in Table 3 and Table 4, and the label distributions are outlined in Table 5.

Task: Predicting Faithfulness of News Representations on Social Media — We propose a new task to predict the faithfulness of news representations on social media by classifying tweets with links to news articles according to subjectivity and information inheritance.

Our proposed system combines RoBERTa [3] encoders and a logistic regression model. The text of the tweet, the title of the article, and the article text are inputs to separate encoders, where the final hidden layers of the [CLS] token for each input text are concatenated and passed to the logistic model, which is trained to predict the correct Subjectivity and Information Inheritance labels.

We conduct 10-fold cross-validation across both the Subjectivity and Information Inheritance datasets. The input Japanese text is tokenized using JUMAN++ (`https://github.com/ku-nlp/jumanpp`) and sent to a pretrained Japanese RoBERTa model (`https://huggingface.co/nlp-waseda/roberta-large-japanese`). Our model is compared to a ZeroR baseline.

Results shown in Table 6 and 7 indicate a significant improvement over the baseline, along with a large potential for improvement, particularly for the information inheritance labels.

Conclusion — In this study, we conducted an analysis of twitter posts containing references to news items, constructed a dataset that captures this tweet-article relationship, and finally, constructed a classifier that solves the novel task of predicting faithfulness of news representations on social media. Future work may focus on increasing the dataset size, accounting for more forms of information alteration, such as translation, and improving the performance of models.

References

[1] J. Thorne and A. Vlachos. Automated fact checking: Task formulations, methods and future directions. In *Proc. of COLING 2018*, pages 3346–3359, August 2018.

[2] C. Samarinas, W. Hsu, and M. Lee. Improving evidence retrieval for automated explainable fact-checking. In *Proc. of NAACL-HLT 2021: Demonstrations*, pages 84–91, June 2021.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.

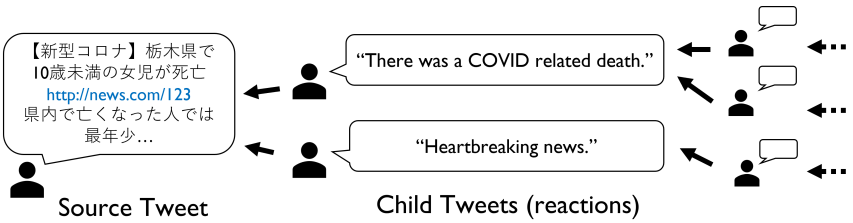

Figure 1: An example of a reaction tree.

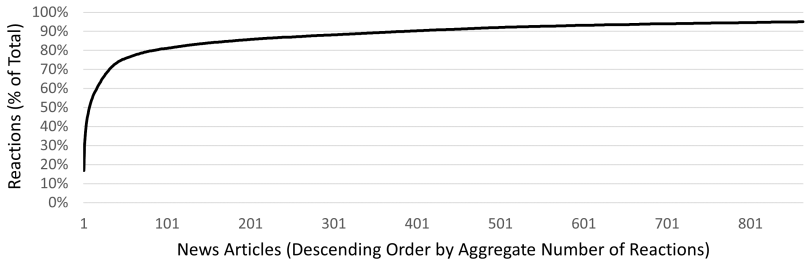

Figure 2: Aggregate twitter reactions for news links posted on May 1. (Top 95%)

Category Label	Tweets	% of Total
Sum. / Paraphrasing	227	24.8
External Reference	122	13.3
Opinion	531	58.1
Title Repetition	745	81.5
Total	914	-

Table 1: Category label distribution.

Alteration Label	Tweets	% of Total
No Alteration	105	46.3
Minor Alteration	100	44.1
Major Alteration	10	4.4
Fabrication	12	5.3
Total	227	-

Table 2: Alteration label distribution for ”summarization/paraphrasing” category.

Label	Criteria
Objective	Doesn’t have ”Opinion” label
Mixed	Has ”Opinion” and other labels
Subjective	Has only the ”Opinion” label

Table 3: Attribution criteria for subjectivity.

Label	Criteria
High	Has "Sum. / Para." with minor alt. or less
Neutral	Has "External Reference"
Low	Has "Sum. / Para." with major alt. or more

Table 4: Attribution criteria for information inheritance.

Subject.	
Label	Tweets
Objective	155
Mixed	184
Subjective	347

Info. Inherit.	
Label	Tweets
High	710
Neutral	120
Low	22

Table 5: Distribution of labels for subjectivity and information inheritance.

System	Precision	Recall	F1
ZeroR	0.17	0.33	0.22
Proposed	0.64	0.62	0.62

Table 6: Model performance for the proposed method (Subjectivity). Scores are macro-averaged.

System	Precision	Recall	F1
ZeroR	0.28	0.33	0.30
Proposed	0.49	0.46	0.46

Table 7: Model performance for the proposed method (Info. Inherit.). Scores are macro-averaged.