1 Data and code access

We will follow all guidelines in the "Redistribution of Twitter content" in Developer Terms. To get access to the data, you can send a request to `rhu@wpi.edu` and we will send you a link to get access to the data repository. To get access to our code, you can navigate to our Github repository: https://github.com/ruofanhu/Tweet-FID

2 Data statement

Tweet-FID includes both the crowdsourced annotation set and the unified expert label set for 4122 unique tweets. These 4122 tweets are sampled from streaming data retrieved through the use of Twitter API with keyword filtering since January 2019. The search keywords are based on food safety literature best practices [1], and include common terms and hashtags that are intuitively indicative of foodborne illness, including "#foodpoisoning", "#stomachache", "food poison", "food poisoning", "stomach", "vomit", "puke", "diarrhea", and "the runs". We also extract about 1000 tweets without mentions of these keywords to diversify the dataset.

2.1 Crowdsourced annotation set

We collect annotations on the crowdsourcing platform Amazon MTurk. Annotators are first asked to label food, location, symptoms, and foodborne illness keywords (*e.g.*, terms like food poisoning, food poison, foodborne illness) in the given tweet. Thereafter, the annotators are to rate the tweet by the degree to which they agree with the statement that this tweet indicates a possible foodborne illness incident (using a Likert scale). Subsequently, assuming the tweet has been indicated as a foodborne illness incident, then for each of the selected entities, the annotator is asked to decide if the entity is related to this foodborne illness incident. We set the restriction that at least one entity must be a relevant entity if and only if the tweet indicates a possible foodborne illness incident.

Each tweet has at least 5 annotations on the three facets described above. The raw collected annotations are stored in CSV files (both the initial batch for the pilot study and batches published in the main study). The processed data, which is grouped and aggregated by **tweet id** is stored in a CSV file.

2.2 Gold standard set

The gold standard set Tweet-FID includes 4122 tweets and 1362 of them are relevant to foodborne illness incidents and 2760 are irrelevant tweets. This set is stored in pickled file format. For each tweet, there are 6 attributes: *id, tweet, tweet tokens, sentence label, entity label, related label, and entity relevance label*. The last three entity labels are corresponding to the tweet tokens. The *entity relevance label* can be derived by a combination of *entity label* and *related label*.

We do a train-validation-test split for Tweet-FID. The training set consists of 1088 relevant tweets and 2210 irrelevant tweets. The validation set includes 137 relevant tweets and 275 irrelevant tweets. The test set consists of 137 relevant tweets and 275 irrelevant tweets. The split ratio is close to 8:1:1. This split is stratified by the tweet-level relevance class. The ratios of relevant to irrelevant tweets in these three splits are the same.

2.3 Usage of the dataset

The usage of the crowdsourced annotation set is to develop more advanced data aggregation methods. For now, we conduct an ad hoc hierarchical aggregation for the three labels, however, the experiment results show that the aggregation result for the entity relevance still has a big room to improve. To the best of our knowledge, there are no interactive aggregation methods that could aggregate the annotations on multiple facets of one object simultaneously.

The aggregated labels only can serve as the weak supervision for the downstream tasks because they are still inaccurate or incomplete. These weak labels can be used to develop advanced single task or multi-task weakly supervised learning models. A more ideal method is to build an end-to-end model which can directly learn the classifiers from the crowdsourced labels.

3 License

The person in request ("the user") may receive and use Tweet-FID ("the dataset") only after accepting and agreeing to both the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy and the following terms and conditions:

Commercial and academic use

The dataset is made available for non-commercial purposes only. Any commercial use of this data is forbidden.

Redistribution

The user is not allowed to copy and distribute the dataset or parts of it to a third party without first obtaining permission from the creators.

Publications

The use of data for illustrative purposes in publications is allowed. Publications include both scientific papers and presentations for scientific/educational purposes.

Citation

All publications reporting on research using this dataset have to acknowledge this by citing the following article:

Hu, Ruofan, Dongyu Zhang, Dandan Tao, Thomas Hartvigsen, Hao Feng, and Elke Rundensteiner. *"Tweet-FID: An Annotated Dataset for Multiple Foodborne Illness Detection Tasks"*, In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6212-6222. 2022.

For specific software output that is shared as part of this data, the user agrees to respect the individual software licenses and use the appropriate citations as mentioned in the documentation of the data.

Tweet-FID changes

The creators of this dataset are allowed to change these terms of use at any time. In this case, users will have to accept and agree to be bound by new terms to keep using the dataset.

Warranty

The dataset comes without any warranty. In no event shall the provider be held responsible for any loss or damage caused by this data.

4 Data management plan

Data archiving, access, and data preservation. Data is stored in Open Science Framework (OSF). The stored data can be protected by this repository. The creators also keep the backups of these files in the server of Worcester Polytechnic Institute(WPI). The stored data will be protected with disk mirroring, daily backups, and other means. Full-time system administrators will monitor the security and availability of these systems. Appropriate access control and other security policies and mechanisms will be put in place to protect the integrity, security, privacy, confidentiality, and other rights or requirements.

Data description and formats. Tweet-FID dataset contains 4126 tweets. The raw collected annotations are stored in CSV files (both the initial batch for the pilot study and batches published in

the main study). For easier access, the gold standard dataset with sentence class label, entity label, relevant entity label, and entity relevance label is stored in both JSON and pickled file formats.

Data privacy. All tweets collected in the dataset are public, but some tweets may be deleted by the user. To minimize the potential influence on the tweeters, we have converted user mention and URL links to `@USER`, `HTTPURL`, respectively, which could to some degree protect user privacy.

Policies and provisions for re-use, re-distribution. We will follow all guidelines in the "Redistribution of Twitter content" in Developer Terms of Twitter. As some tweets may be deleted by its users, in the future, researchers may not be able to get access to all Twitter content in such a case. For this reason, we will redistribute the hydrated Twitter content to assure the value of this curated data set is preserved. This falls within the guidelines of Twitter because our current dataset is smaller than 50,000 tweets. Furthermore, before we give access to researchers on this hydrated data set, we will require them to send a request to us and agree to the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy.

Under the current license terms, the dataset is made available for non-commercial purposes only, any commercial use of this data is forbidden. The user is not allowed to copy and distribute the dataset or parts of it to a third party without first obtaining permission from the creator.

The creators of this dataset are allowed to change these policies and provisions for use at any time. In this case, users will have to accept and agree to be bound by new terms to keep using the dataset.

Rights and obligations. The creators of this dataset own the copyright of this dataset. For specific software output that is shared as part of this data, the user agrees to respect the individual software licenses and use the appropriate citations as mentioned in the documentation of the data.

5 Author statement

This is an author statement ("this statement") regarding the paper entitled: Tweet-FID: An Annotated Dataset for Multiple Foodborne Illness Detection Tasks ("the article") in Proceedings of the Thirteenth Language Resources and Evaluation Conference 2022.

All authors (collectively "we") of this article are:

Ruofan Hu, Dongyu Zhang, Dandan Tao, Thomas Hartvigsen, Hao Feng, Elke A. Rundensteiner

We hereby grant a custom license for the dataset Tweet-FID to the general public, which is specified at Tweet-FID Dataset License (available in supplementary materials). We hereby grant an MIT license in the code of this article to the general public, which is specified at `https://opensource.org/licenses/MIT`.

We warrant that:

1. The article is original, has not been formally published in any other peer-reviewed journal or a book or edited collection, and is not under consideration for any such publication.

2. We are the sole author(s) of the article, and we have a complete and unencumbered right to make the grants we make.

3. The article does not libel anyone, invade anyone's copyright, or otherwise violate any statutory or common-law rights of anyone, and we have made all reasonable efforts to ensure the accuracy of any factual information contained in the article.

References

[1] Thomas Effland, Anna Lawson, Sharon Balter, Katelynn Devinney, Vasudha Reddy, HaeNa Waechter, Luis Gravano, and Daniel Hsu. Discovering foodborne illness in online restaurant reviews. *Journal of the American Medical Informatics Association*, 25(12):1586–1592, 2018.