An Evaluation Benchmark for Online Discussion Representation Models

Anonymous Author(s)
Affiliation
Address
email

1 Motivation

As access to digital media becomes easier and more commonplace, the volume of communications over Web platforms increases. Great volumes of data are generated every moment,[1] and all this information can be explored for various objectives. Search engines and recommendation systems are always present in online environments, and many other more specific tasks are solved with the help of data obtained online.

One of these sources of content is the massive amount of comments written by users in various websites. The active participation of commenters in discussion sections is such that 46% of social network users have already participated in at least one discussion in news posts (Anderson and Caumont, 2014). With so many users generating data all the time, spontaneously and costlessly, online discussions prove to be valuable sources of data for researches and other interested parties. Despite the fact that dealing with comments brings a series of challenges (such as the informality of the language used in them, the constantly changing vocabulary, and the question of legitimacy regarding who generated certain comments), online discussions are still employed in several research works (Tigunova et al., 2019; Cheng et al., 2019; Hoogeveen et al., 2018).

However, each works targets a different problem, often using unique datasets. Researchers propose novel representations for discussions or comments: feature sets, embeddings, and distributed vectors. With each work aiming at a different objective, it becomes difficult to know how well a certain representation model performs outside of the tasks it was tested on, and the utility of a proposed model becomes limited to the work it was first intended for.

Bhatia et al. (2014), for example, use both textual features and dialogue act labels for extractive summarization of discussion threads, using comments from the official *Ubuntu Linux* distribution forum and the *Trip Avisor* forum. Considering another task, Wang et al. (2012) use discourse structure features to classify the *solvedness* of a thread, experimenting on threads crawled from *Linuxquestions* and *Debian mailing lists*. Meanwhile, Kano et al. (2018) use neural models to extract content and context features for the same task of summarization, but using *Reddit* threads. Also using *Reddit* discussions, Kumar et al. (2018) use lexical and stylistic-linguistic features to classify the sentiment of the source post of each thread. As a final example, the work of Backstrom et al. (2013) uses data from *Facebook* and *Wikipedia* discussions to predict thread length and return of participants, employing textual features, as well as features regarding time of comments, user IDs, presence of hyperlinks, and so on.

2 Goals and Contributions

In order to make it easier to compare works dealing with online discussions, and to better figure out how well specific models function outside of their intended domains, we propose a benchmark

[1] `https://www.webfx.com/internet-real-time/`

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

to evaluate online discussion representation models. This benchmark will provide a collection of discussion datasets, a set of evaluation metrics for different tasks, and some baseline representation models.

2.1 Discussion datasets

Firstly, we release a collection of datasets containing online discussions and related data. Each of these datasets has comments from a different web forum, usually dealing with completely different domains. They also have varied associated characteristics, allowing each of them to be evaluated in different tasks. Currently, the following datasets are being explored:

- RShows : Comments from `Reddit`, containing episodic discussions regarding series of different genres;
- YT8M : YouTube comments for videos taken from the *YouTube-8M* dataset Abu-El-Haija et al. (2016);
- MAL : Comments from the `MyAnimeList.net` forum, containing episodic discussions regarding animated television series;
- GameF : Comments from the `GameFAQs` forum, containing discussions regarding video games titles;
- GReads : Comments from the `Goodreads` forum, containing discussions regarding books.

We also explore previously published discussion datasets, such as the *New York Times Comments* dataset (Kesarwani, 2018), containing discussion sections from New York Times articles, and the *Yahoo News Annotated Comments Corpus* dataset (Napoles et al., 2017), containing discussion sections from Yahoo News articles.

2.2 Evaluation tasks

Secondly, we propose a series of evaluation tasks for discussion representations, each with their own metrics of quality. Some of these tasks can only be performed in some of the datasets, as certain characteristics vary from one domain to another. The following tasks have been defined so far:

- TClust : Attempting to cluster discussion threads according to their representations, checking if threads clustered together belong to the same subject;
- TOrder : Comparing the order defined by the proximity between threads to an external order the discussions should follow;
- SRecom : Item recommendation according to how close one discussion representation is to others;
- CSelect : Selection of the most representative comments from each discussion.

2.3 Representation models

Finally, we evaluate an initial selection of representation models on the datasets, with room for additional models being implemented later, according to the proposed tasks. The representations to be tested are based on the following methods:

- TFIDF : Representing each discussion as a simple *TF-IDF* vector, treating the entire discussion as a document;
- TKT : Representing each discussion as a *TF-IDF* vector that considers only the *Top-k Terms* from the dataset;
- doc2vec : Learning distributed representations for each discussion according to the *Paragraph Vector* methods from Le and Mikolov (2014);
- DeepN : A novel deep neural network model for discussion representation.

The code for the complete benchmark will be publicly available, as well as the datasets used in this work.

References

Abu-El-Haija, S., N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan 2016. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675.

Anderson, M. and A. Caumont 2014. How social media is reshaping news. `http://pewrsr.ch/1tZ2Rsu`. [Online; accessed 10-September-2019].

Backstrom, L., J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil 2013. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, Pp. 13–22, New York, NY, USA. ACM.

Bhatia, S., P. Biyani, and P. Mitra 2014. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pp. 2127–2131, Doha, Qatar. Association for Computational Linguistics.

Cheng, H., H. Fang, and M. Ostendorf 2019. A dynamic speaker model for conversational interactions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Pp. 2772–2785, Minneapolis, Minnesota. Association for Computational Linguistics.

Hoogeveen, D., L. Wang, T. Baldwin, and K. M. Verspoor 2018. Web forum retrieval and text analytics: A survey. *Found. Trends Inf. Retr.*, 12(1):1–163.

Kano, R., Y. Miura, M. Taniguchi, Y.-Y. Chen, F. Chen, and T. Ohkuma 2018. Harnessing popularity in social media for extractive summarization of online conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kesarwani, A. 2018. New York Times Comments. `https://www.kaggle.com/aashita/nyt-comments`. [Online; accessed 18-September-2019].

Kumar, S., W. L. Hamilton, J. Leskovec, and D. Jurafsky 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press.

Le, Q. and T. Mikolov 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, Pp. 1188–1196.

Napoles, C., J. Tetreault, A. Pappu, E. Rosato, and B. Provenzale 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*. Association for Computational Linguistics.

Tigunova, A., A. Yates, P. Mirza, and G. Weikum 2019. Listening between the lines: Learning personal attributes from conversations. In *The World Wide Web Conference on - WWW '19*. ACM Press.

Wang, L., S. N. Kim, and T. Baldwin 2012. The utility of discourse structure in identifying resolved threads in technical user forums. In *Proceedings of COLING 2012*, Pp. 2739–2756, Mumbai, India. The COLING 2012 Organizing Committee.