Data statement of stance-annotated Danish Reddit dataset

Anders E. Lillie and Emil R. Middelboe

June 2, 2019

1 Introduction

This is the data statement [Bender and Friedman, 2018] for a stance-annotated Reddit[1] dataset for the Danish language, denoted DAST.

2 Curation rationale

Reddit Submissions and user posts are the main content of the DAST dataset. Furthermore, the data includes user information such as account creation date, Reddit karma, gold status, if the user is a Reddit employee, and whether the user has a verified e-mail address. Finally the user ID is included, which is unique for each user.

For each submission the following is included: submission ID, title, text (if any), creation date, number of replying comments, direct URL reference, a text URL (if any), number of upvotes, and whether the submission post is a video. Finally, annotations following [Zubiaga et al., 2016] of the following are included: if the submission initiates a rumour, and if if this is the case, its truth status, as well as a "Supporting", "Denying", or "Underspecified" label towards the rumour for the submission post, as well as SDQC annotation for each replying posts (and nested replies).

Each individual comment contains the following: comment ID, text, ID of the post replied to, direct URL to the comment, its creation date, upvote/downvote score, whether the user of the comment is the submitter of the submission, and the number of replies. Furthermore certainty and evidentiality annotations are included [Zubiaga et al., 2016].

The DAST data primarily contains text from the Danish language. However references and words from the English language occurs from the Danish users in parts of the text.

3 Speaker demographic

The speaker demographic is unknown since the dataset is from an anonymous platform.

[1] `https://www.reddit.com/`

4 Annotator demographic

Both annotators are male Danish students, respectively 24 and 25 of age, on the Software Development master programme, at the IT University of Copenhagen. Neither have any training in linguistics or annotation processes, but have studied the field of Stance NLP and machine learning. Further, both students graduate as MSc students in June, 2019, with a cand.it degree.

5 Speech situation

The dataset contains Reddit submissions ranging from 2012 to 2019. It consists only of written text and no sign/spoken language is used. The interactions are asynchronous.

6 Text characteristics

The text is from the social media/microblog website Reddit, spanning over a number of different topics, which are introduced in the next section.

7 Other

This section presents the volume of DAST including an overview of the data events and SDQC annotations. First, the data is presented in table 1.

Event	*Submissions*	*Branches*	*Posts*
5G	4	117	273
Donald Trump	3	89	246
HPV vaccine	7	122	255
ISIS	2	68	169
"Kost"(diet)	3	165	557
MeToo	1	29	60
"Overvågning"(surveillance)	1	121	352
Peter Madsen	3	156	381
"Politik"(politics)	3	126	323
Togstrejke(train strike)	2	49	101
"Ulve i DK"(wolves in DK)	4	119	290
Total	33	1,161	3,007

Table 1: Overview of data events and submissions

In total, DAST contains 3,007 Reddit posts, which are distributed across 33 Reddit Submissions respectively grouped into 16 (Danish labelled) events. DAST is also annotated for stance using the SDQC annotation scheme from [Zubiaga et al., 2016]. The SDQC distributions for respectively annotation targeted towards the rumour and replying posts (parent posts) are illustrated in table 2.

Finally the dataset was also annotated for rumours, with a total of 16 submissions deemed as rumours. Out of the 16 rumourous submissions, three of them were true, three were false and the rest were unverified. In total, they

Label / *Target*	S	D	Q	C
Reddit Submission post	273	300	81	2,353
Reddit parent comment	261	632	304	1,810
Reddit Submission post %	9.1	10	2.7	78.2
Reddit parent comment %	8.7	21	10.1	60.2

Table 2: Relative SDQC stance label distribution for DAST

make up 220 Reddit conversations, equivalent to 596 branches, and a total of 1,489 posts.

The posts are distributed across nine events as follows: 5G(233), Donald Trump(140), ISIS(169), "Kost"(324), MeToo(60), Peter Madsen(381), "Politik"(49), "Togstrejke"(73), and "Ulve i DK"(56). Thus ISIS, MeToo, and Peter Madsen are the only events which only contain rumourous conversations.

References

[Bender and Friedman, 2018] Bender, M., E. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science.

[Zubiaga et al., 2016] Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., and Tolmie, P. (2016). Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS ONE*. 11(3).