Problem Statement:

To do sentiment analysis and come up with a classification algorithm to identify whether post contains cyberbullying content or not.

Dataset Description:

1. The data represents 50 ids from Formspring.me that were crawled in Summer 2010. For each id, the profile information and each post (question and answer) was extracted. Each post was loaded into Amazon's mechanical turk and labeled by three workers for cyberbullying content.
2. The data contains the following profile fields:
    a. BIO - profile biography created by owner of the id
    b. DATE - the date the id was crawled
    c. LOCATION - location provided by the owner of the id
    d. USERID - The actual id itself
3. The data contains the following information on each post
    a. TEXT - the question and answer (separated by Q: and A:)
    b. ASKER - the id of the person asking the question (blank if anonymous)
4. The occurrences of label data:
    a. ANSWER - YES or NO as to whether the post contains cyberbullying
    b. CYBERBULLYINGWORK - word(s) or phrase(s) identified by the mechanical turk worker as the reason it was tagged as cyberbullying (n/a or blank if no cyberbullying detected)
    c. SEVERITY - cyberbullying severity from 0 (no bullying) to 10
    d. OTHER - other comments from the mechanical turk worker
    e. WORKTIME - time needed to label the post (in seconds)
    f. WORKER - mechanical turk worker id

Download the kaggle dataset here.

NOTE: Do have a look at an example notebook here to gain better understanding of the task.