Do Computer Science Scholars Consider Issues of Privacy when Studying Large Twitter Data Sets?

@ShirleyEarley
@melissaterras
@clhw1

Shirley A. Williams[1], Melissa Terras[2], Claire Warwick[2]
[1]School of Systems Engineering, University of Reading, UK
[2]Department of Information Studies,
University College London, UK

What is Twitter?

What do researchers do with Twitter?

Launched in 2006;
Three academic papers in 2007;
… Thousands in 2011

@SciVerse_Scopus
How many Computer Science mention Twitter in their abstracts?

1058

Random Sample = 282
Twitter-focussed = 162

Primarily about Twitter as opposed to just mentioning it.
None were off-topic – unlike studies of other corpus

We categorise large scale Twitter studies as those considering over a million tweets and/or a million accounts

Williams S, Terras M, Warwick C (2013) “What people study when they study Twitter: Classifying Twitter related academic papers” Journal of Documentation 69: in press.

Read and re-read – gives 55 papers that are are large studies

Few give details of data in the abstract.
Some that obviously Twitter data don’t even give details in the paper!

26 make some consideration of privacy

29 do not

For example explaining they collect from the public time-line

No mention of the Twitter privacy policy

https://twitter.com/privacy

Some misunderstand what protected means in Twitter

“Accounts with protected Tweets require manual approval of each and every person who may view that account's Tweets.” *Twitter support*

The account details are not protected….

What are they studying Twitter data to find out?

Derived characteristics; Classifying on personal information; Location; Spam; other

Great east Japan earthquake viewed from a URL shortener

They made their dataset available and the paper discusses issues of privacy and making data anonymous.

Tweeting about the tsunami? - Mining Twitter for information on the Tohoku earthquake and tsunami

No consideration of privacy

Crowd-based urban characterization: Extracting crowd behavioral patterns in urban areas from Twitter

No consideration of privacy

Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis

No consideration of privacy – and real tweets used as examples (but not ids)

The majority of Computer Science scholars do not consider ethical issues in relationship to privacy when studying large Twitter data sets.

- Twitter users need to be more aware of who sees what
- Researchers need to be aware of the ethics of handling Twitter data.

ごせいちょう
ありがとう ございます
Thank you for listening
from @ShirleyEarley

Further work – Twitter research and beyond