

Report on the IPS Digital Frontiers Seminar: “Assessing the Rationality of Political Online Space: Man and Machine”

11 February 2015

By Nanditha Das
IPS Research Associate

The online space has been described as a “Wild, Wild West” rife with Distortions, Rumours, Untruths, Misinformation and Smears (“DRUMS”, coined by a Member of Parliament). While the Internet has been used to galvanise people for the collective good, many have voiced concerns over the noise and vitriol present online, cyber harassment and polarisation.

At a seminar held on 11 February 2015, the Institute of Policy Studies (IPS) researchers Dr Carol Soon and Mr Tan Tarn How presented their study of 190 over blogs with social and political content during the period of June–July 2014. They analysed the blogs for their structural characteristics and “rationality”. In the second part of the seminar, Professor Lim Ee-Peng from the Living Analytics Research Centre (LARC), Singapore Management University, presented initial results of large-scale and real-time content labelling by machine of the same IPS data set.

The seminar was attended by over 50 participants, including a Member of Parliament, policymakers, academics, bloggers, members of the media and representatives from community organisations.

I. Opening Remarks by IPS Senior Research Fellow Tarn Tan How

Mr Tan said the seminar is the first of IPS’ new Digital Frontiers Series, based on the idea that new technologies have opened the door for novel ways of doing things in the social, political, economic and research arenas.

Mr Tan informed that the second seminar, to be held in mid-2015, will be on the “Digital Village”. For that, IPS Research Fellow Dr Carol Soon will present a conceptual paper on Singapore as a digital village of six million people. The third will be a seminar by Mr Arun Mahizhnan, IPS Special Research Adviser, on “Gov.Com: 1 to 1”, addressing how advancements in digital media technology has revolutionised the way in which the government interacts with citizens, and citizens with other citizens. He will explore the ways in which governments and citizens can take advantage of the new opportunities afforded by digital technologies.

Turning to the present, Mr Tan started by asking if rationality should even be expected of the online space. The Internet is very different from traditional media, and requirements for rationality as expected of traditional media may not apply. Research has established that the motivations which drive online behaviour are typically focused on the self, e.g. self-expression, self-promotion and social networking. Blogs and social media are popular because they offer the person on the street a free and easy, and even visceral and emotive way to communicate with others. Hence the question is whether the online space should be held to the same benchmarks as traditional media for fact-reporting as they serve different purposes.

He also introduced the Political Online Space Sensing Project that IPS is spearheading. The project involves scholars from various institutions and disciplines using different methodologies to study the online space. Mr Tan highlighted the challenges of studying public opinion due to limitations of existing measurement tools. Many studies of surveys have shown for instance that people tend to give socially acceptable answers, fear reprisals for telling the truth, or fail to recall accurately their own actions. Social media analysis also faces issues such as the representativeness of the users of the larger population, the difficulty of getting basic information such as the sex, age and social economic status of users, and the rapidly shifting profiles of users and patterns of use. Hence, it is important to be aware of the potential pitfalls of each methodology and mitigate them as far as possible. The multi-stakeholder collaboration will leverage the strengths of traditional methods (i.e., content analysis and survey) and new big data techniques for sensing, tracking and predicting public opinion. In addition, the collaboration will enable the comparison of different tools and identify ways in which they can be used to better understand people's attitudes and opinions in cyberspace, and the convergence and divergence of opinions between the offline and online spaces.

II. Presentations

The two presentations made at the seminar are summarised below.

1) ***"Study on the Rationality of the Political Online Space" by Dr Carol Soon, Research Fellow, Institute of Policy Studies***

Dr Soon discussed prevailing concerns about the Internet's polarising effects, how it may divide real world communities and the consequences of various forms of "irrational" behaviour such as cyber lynching, cyber vigilantism and corrosive speech.

Dr Soon raised the question of what "rationality" means. Theoretical scholars such as Alasdair Macintyre and Alberto Guerreiro Ramos had argued that there is no universal definition of rationality; changing political and cultural contexts shape the definitions of rationality.

The concept of rationality as adopted in the IPS study is based on traits and values ascribed to it in official and mainstream media discourse. The study measured rationality using three measures: objectivity, emotionality and partisanship. It adopted existing definitions for objectivity, emotionality and partisanship in journalism and political communication research.

She stressed that the study established what is present in the online space, and not which parts of it has the most impact.

The researchers conducted a content analysis on blogs, which had at least one social or political post in the months of June and July 2014 — months that saw hot-button issues such as the National Library Board book controversy, the Pink Dot and Wear White campaigns and the Central Provident Fund (CPF) debates.

The researchers excluded blogs by non-governmental organisations, political organisations and e-versions of traditional media in this study. A total of 197 blogs were identified for the study and researchers analysed more than 1,000 posts. Content analysis was used to determine structural characteristics such as how political a blog was, its age and identity of the blogger (i.e., known or unknown); rationality (i.e., objectivity and emotionality); and partisanship (i.e., for both government and the opposition). The researchers also tested for correlations between bloggers' identity, blog type (i.e., how political a blog is), objectivity, emotionality and partisanship.

(Please refer to the [presentation slides](#) for the coding scheme, samples of different types of posts and detailed findings.)

The findings showed that political issues were discussed by both “highly political” blogs as well as blogs that were mostly about non-political subjects. This suggested that a wide range of people were responding to what was happening offline. Bloggers who mentioned the government in their posts were somewhat or very anti-government, while those who mentioned the opposition tend to be somewhat or very pro-opposition. The findings also confirmed the presence of anonymous “keyboard warriors” who blogged about political issues, raising the question of whether there are fears of repercussion for doing so publicly.

Dr Soon concluded that the political online space in Singapore was in fact not the “Wild, Wild West” as feared. About 50% of the bloggers were somewhat or very calm in their posts. Almost 70% presented an alternative side — sometimes multiple sides — to an issue. “Highly political blogs” tended to be anti-government but a good number were also critical of the opposition or pro-PAP. While there were exceptions, “highly political” blogs tended to be objective regardless of their partisanship, suggesting perhaps that bloggers recognise that balance help to sway minds and change opinions.

Mr Tan discussed the limitations of the study. First, it was a snapshot of something much larger: it covered only June and July 2014 and left out other social media such as Facebook and Twitter. Facebook was left out because strict privacy settings made harvesting data difficult. Twitter’s 140-character limit made it difficult to analyse content for rationality.

Mr Tan concluded the presentation by discussing possibilities for future research. The Political Online Space Sensing Project will attempt to do what he called “triangulation” — between different ways of measuring public opinion, from surveys to elections to social media analysis. For example, IPS has three terabytes of data archived from Facebook, Twitter, YouTube, mainstream and alternative media. He suggested such information could be compared with the traditional survey data and the election results to develop a model that can predict public opinion for the next election.

2) “Possibilities of Machine Classification” by Professor Lim Ee-Peng, Director, Living Analytics Research Centre (LARC), School of Information Systems, Singapore Management University

Professor Lim explained how computers were used to classify or code online content. First, researchers decide the kind of information they want to extract from the content (for this study, it was in the form of text, but it can be photos, videos, etc.). Examples are “topics”, “sentiments”, or “party orientation”. Second, for each type of information, the researcher determines the categories or labels under which the computer should classify each piece of content. Labels under “topics” could be CPF, housing, immigration. “Sentiment” might be labelled as positive, neutral or negative. “Party orientation” might have labels like pro-government, anti-government, pro-opposition or anti-opposition. One method is for the researcher to have a set of keywords for each label that the machine will locate. The machine then looks for these keywords in each piece of text to decide its category. A second method is for the researcher to tell the computer what label each text falls under, so it can derive a set of keywords for each label.

The computer’s prediction results are compared to the classification done by humans to see if the computer is as accurate. The machine can be given more data from human classification to improve its accuracy.

For humans to classify texts accurately, they need to be familiar with the context and be trained to do so. The human approach is laborious and hence not scalable. With the exponential growth of data, human classification of everything is not possible. The question arises to what extent machines can help. Even though machines may not perform as accurately as humans, they can be trained quickly to classify much larger amounts of data than humans. A trained machine classifier that is integrated with a data crawler (i.e., a programme that harvests the massive amount of online information) can perform large-scale label prediction quickly.

There are still some technical challenges in developing machine classifiers for online content, Prof. Lim said. In data crawling, for example, programmers have to cope with different online content formats. Specialised crawlers for websites and blogs therefore have to be developed. Social media sites also impose restrictions on crawling. Facebook, in particular, is difficult to crawl compared with Twitter and Instagram. The second issue is the heterogeneous nature of data. For example, Twitter messages are very short. Although content is easily accessible, it may not be easily understood.

The other challenges include “noise” in the form of advertisements embedded in online content, which has to be removed before analysis; the difficulty of analysing multi-page content; and the dynamic nature of online data. People often write ungrammatically, misspell words or use social media lingo.

Prof. Lim said that in his study, topic labels identified with IPS researchers included “censorship”, “CPF”, “foreigners”, “inequality” and “LGBT”. LARC researchers then classified blog posts based on these labels. Using a corpus of 5,721 blog posts, LARC researchers conducted cross-validation and logistic regression. The results were reasonably accurate. In

addition, the classifier learning process also identified new relevant keywords that had not surfaced earlier.

Then, the machines classified blogs for political orientation, e.g., “pro-government” and “anti-government”. The prediction for anti-PAP posts was more accurate than that for pro-PAP posts.

Prof. Lim emphasised the need to take text classification beyond experiments and turn it into systems that demonstrate the capability of classifying large content. He highlighted the possibilities of scalability in classification and working with real-time data when coding is done in conjunction with classification. This method does not disturb the end users, unlike traditional surveys. For example, it is possible to conduct real-time crawling and classification of blog posts, and detect emerging topics during and after the elections.

For future research, Prof. Lim said it would be interesting to monitor lurkers, i.e., users who observe but do not participate, and the types of content they access. User attention analysis is another area for further study. Prof. Lim concluded his presentation by suggesting the possibility of turning text analytics into games that would channel more human effort to labelling online data.

III. Discussion

Next, Mr Tan moderated the discussion. The following issues were raised.

1) *Sampling of blogs*

A participant asked the researchers how they defined a political blog. Dr Soon explained that the researchers adopted definitions used in existing research. Political blogs are blogs that discuss a broad range of social, economic, political and governance issues. They include the government, policies, Members of Parliament, political parties, and policy issues on transportation, education, religion and censorship.

2) *User attention analysis*

Another participant highlighted the issue of user attention, in particular, where readers were spending their time online. He said that it would be interesting to obtain details on user attention and study the more rational blogs that attract a higher readership.

Prof. Lim agreed that there is a gap and said there is a need to conduct more fine-grained analyses of user behaviour in order to quantify user attention. One way would be to look at the number of “likes” or “shares” by users. He said it is important to interpret data based on whether they originate from individuals, small groups or larger communities.

Mr Tan added that future studies could also include other indicators such as the number of comments and unique visits. He stressed the need to analyse the readership, its purpose and effects.

3) Analysis of interactions across different platforms

Another participant asked whether the impact of user interaction could be analysed, citing the example of Facebook, which has pages ranked based on “likes” and “shares”. Can a cross-comparison of what users say on their blogs, Facebook pages and Twitter be done? He noted that there might be a link if one pays attention to the user consistency across all the social media platforms. Prof. Lim clarified that they did not analyse “likes” on face-value as “likes” may have other meanings beyond endorsement, and there is a possibility of manipulation. Hence, “likes” may not reflect online sentiments accurately. His research group focuses more on user attention by topic and using this principle to determine influential users on Twitter. Their research results have been published and are open to the public.

Mr Tan said that Prof Lim’s study classified individual blog posts whereas the IPS study classified blogs as a whole. He also highlighted the difference between studies of existence (i.e., what is out there) and of impact (i.e., what has the most traction or readership). It is important to analyse the impact but this study was not about that, but more about existence. Thus, this study found that there were a number of pro-government and neutral blogs, but did not examine if they had any impact or why.

4) Feedback to policymakers

The same participant sought clarification on whether research findings are shared with policymakers and if Internet Protocol (IP) addresses are monitored. Prof. Lim said that they did not collect information on IP addresses.

Another participant shared her experience: where government is interested in research in this domain and input goes to the government via academic publications that are publicly available. She also acknowledged that the output may not be what the government wants to hear sometimes but it could be due to differences in study objectives.

Mr Tan added that the data would be available online. He said that IPS shared the raw data of their 2011 survey with the government and plans to make the information available to the public. Sharing data with the government helps it be more responsive to citizens’ feelings and formulate better policies.

5) Framing of messages

Another participant asked if there was a contradiction for the conclusion that “highly political blogs” are more partisan (i.e., anti-government and anti-opposition) while also being objective. Mr Tan replied that this was indeed a counter-intuitive conclusion and it illustrates the varied dimensions of rationality. Thus, a blog can be for or against the government and still be balanced in presenting both sides of an argument and eventually coming down on one side. It can also present its conclusion for one side in a calm, unemotional way. He asked the audience to consider the case of the mainstream media here, which many consider both partisan and largely balanced and calm.

Dr Soon added that much research in psychology, cognitive science, marketing and public communications show that message presentation affects attitude formation. Two-sided

messages (e.g., messages that present the pros and cons of an issue) have a greater positive effect on attitude change than one-sided messages (e.g., messages that present only the pros of an issue). She explained why the results of the IPS study were not contradictory using an example: if a blogger approves of the government policy on CPF, he or she would present the positive side but may also highlight people's criticisms concerning the CPF policy before concluding the post by supporting the government's position. In this illustration, the blogger is objective (by presenting more than one side of the argument) but is partisan towards the government. Dr Carol shared that if funding becomes available, her new project would examine the effects of information presentation on individuals' attitudes.

Another participant shared that a study she conducted of the 2011 General Election showed that two-sided messaging increased both people's understanding of an issue and their confidence in their stance on the issue. It also makes people more critical concerning the source of information they received, the producer of the information and the context in which the information was presented.

Prof. Lim acknowledged the difficulties in classifying pro-PAP posts. He said it could be due to the subtlety. Although some training data was available, it was limited because of the paucity of pro-PAP posts. For analytical purposes, his study was based on the assumption that certain types of posts were anti-PAP and other types were pro-PAP. He said there is a need to re-examine the assumption and adequate labels can be generated to train the machines with the help of domain experts.

6) Interpretation of online sentiments

Another participant asked why the study was confined to blogs and not to other types of online chatter. She asked the researchers to clarify what they meant by being anti-government. She said a blogger could have a pro-government leaning but could be speaking against the government due to a difference in opinion on a subject or the blogger might even want to spur the government to do better. The participant also asked how the researchers differentiated whether the blogger was critical of the policy or of the institution.

Mr Tan replied that with content analysis, researchers cannot determine bloggers' hidden or actual beliefs by second-guessing, but can only deduce from the contents of their writing. Dr Soon acknowledged that every methodology has its limitations, and content analysis cannot capture bloggers' hidden or unstated intentions or motivations. However, it is a very useful tool to assess the subject matter. The study did not use binary measures (i.e., either pro- or anti-government) but scales with many points (that is, different degrees of pro- or anti-government) to capture nuances in opinions and compensate for methodological limitations.

Prof. Lim explained that blogs are the most accessible resources available publicly. He had faced many issues with Facebook due to the platform's privacy settings. With Twitter, there is insufficient data to draw meaningful conclusions.

On the suggestion of analysing online chatter from more sources, Mr Tan said it was possible with funding support. He also added that blogs were chosen as the units of analysis because their posts tend to be long, and it is easier to draw conclusions on rationality from

longer texts. Another participant agreed that measuring rationality is a complicated process and it cannot be measured in the 140 characters found in a tweet.

7) Future trajectory

A participant asked if available data could be used to gauge the sentiments and issues for upcoming elections. He cited the example of Nate Silver, a statistician who predicted results for the 2008 U.S. presidential election. The participant felt that the majority of online chatter yields less than useful outcomes in the local context, and further analysis is required to provide a more accurate reflection of people's perceptions.

Mr Tan said such research is at an exploratory stage. Forecasting election results is not that hard because one could rely on surveys. It is harder to do so based on social media content. He cited a study of the 2009 German elections, which showed that the frequency of mentions of a party on Twitter accurately predicted how well it did eventually in the polls. Actual tweet content was irrelevant to a party's victory, although tweet content accurately predicted which parties would join as coalition partners after the election. Social media content analysis might show what surveys cannot, such as giving a deeper understanding of what people were thinking about and why, and perhaps how that influenced their behaviour during the polls. This is one aim of the Sensing the Political Online Space Project.

8) Understanding the sentiments

A participant asked whether a blog stance could be targeted at a policy rather than at a party. The findings showed that there was more negativity towards PAP than the opposition. Since the PAP government and not the opposition executes policies, the results could be skewed.

Dr Soon clarified that for many of the issues that emerged from data collection, e.g., Pink Dot, the posts were either neutral or did not mention government at all. She acknowledged that it was not always clear whether people looked at an issue from a policy angle or an execution angle.

Mr Tan added the online space was more conducive for alternative opinions as this space was not available in the mainstream media. He also referred to the finding where the 39% of blogs does not mention "government" at all even though they are political. The researchers cannot read into the underlying stance regarding the government since the study is based on manifest content.

9) Language used in blogging

An audience member commented that machine labelling and the data-mining process is a "smart nation" way to go, but he noticed that all the keywords and labels used in the study were in English. He was curious if there could be a difference in findings for English-language blogs and blogs written in other languages.

Dr Soon referred to an earlier IPS study on Chinese-, Malay- and Tamil-language blogs. Even though she could not make a comparison pertaining to rationality as that study was based on a different research agenda, there were some differences between English- and Chinese-language blogs. Only a small minority of Chinese-language bloggers wrote on

political issues. Most of the topics covered include the Chinese language and culture, and general topics such as travel, food and photography. Chinese-language bloggers who wrote about political issues did so in a very different manner from English-language bloggers. Their commentaries were much more nuanced and this could be due to the nature of the Chinese language and culture, or perhaps political history.

IPS Special Research Adviser Arun Mahizhnan, who focused on the Tamil-language part of the aforementioned study, said that Tamil-language blogs were written by recent immigrants rather than Singaporeans. Apart from this interesting difference, Indian immigrants did not blog about politics much, but when they did, they focused on Indian politics. He suggested that this could be due to their fear of repercussions from engaging in the local affairs. However, he could not draw generalisations as the sample was small. The language used in Tamil-language blogs was also civil and non-abusive.

10) Identification of keywords and topics

Another participant asked about Prof. Lim's study and the keywords he used to decide if a blog post was about a certain topic. Some keywords under some topics had nothing intrinsically linked with the topic but were due to one-off events. For example, the keyword "YMCA" came under the "CPF" topic, but it was due to a one-off event at Hong Lim Park rather than being directly related to CPF. Going forward, future research may not yield accurate findings if such keywords were used to classify blog topics. He asked if Prof. Lim's methodology takes into account such one-off events or keywords.

Prof. Lim said he recognised the limitation and said that keywords used to identify each topic could be continuously revised for correct classification. The participant also asked whether a longer period of study would yield more useful findings than a two-month study. Prof. Lim clarified that the content that was researched was not confined to those written in the two months but covered 5,000 blog posts written over a longer period. For more accurate results, LARC will continue to build their corpus and train the classifier with recent data.

Nanditha Das is a Research Associate with the Arts, Culture and Media Cluster in IPS.

If you have comments or feedback, please email ips.eneews@nus.edu.sg



© Copyright 2015 National University of Singapore. All Rights Reserved.

You are welcome to reproduce this material for non-commercial purposes and please ensure you cite the source when doing so.