Interview: Participant-10

**I: As I was saying, we're looking at tools that people are using for capturing and analysing social media data. We're also interested in how people are linking social media So I'll start off by asking: What discipline or subject does your research fall under?**

M: It's mostly Information Science.. Now, with the social media aspect of it, I guess really it's maybe over the last five or six years or so that they've started looking at how political parties, individual parliamentary candidates, how the official campaigns and the individual parties use social media. I guess both to provide information and to engage and encourage dialogue with the public at large. So I guess that's been the main basis of my studies of the use of social media.

**I: But then you can get some engagement…**

M: That's what we hope, although that hasn’t proved so successful so far in terms of people – people will 'like' photographs or they'll share them but they've been a bit unwilling to give us their own thoughts or memories so far.

**I: That's very interesting; the use of social media as a tool.**

M: I guess the bulk of it has been how political actors, maybe being the broad term used, use social media. But since the emergence of Facebook and Twitter in particular over the last five, six years or so…

**I: There's suddenly a lot more data…**

M: Yeah, absolutely and…

**I: So would you consider your research within the field of Social Sciences?**

M: Yes, very much so. I guess it's people, we're looking to see how the public, the people's representatives or people who are attempting to be their elected representatives are using social media so I think it falls very much within the Social Science field, yeah.

**I: And was that successful in terms of getting the – you mentioned the sort of – you are not getting as much content with the Facebook page presently but when you were engaging politicians directly on policy issues, would you get kind of like a personal response or would you get like a standard…printed letter kind of thing?**

M: It could be a real mixed bag to be honest. Maybe over the course of however many years we've been doing this maybe about half of our questions have been ignored completely. The ones that do respond, again, it can differ. They range really from the copy and paste approach where you will find, particularly with the larger parties- they will just copy and paste chunks of text from maybe their manifesto, other policy documents and just plant them down in the middle of your reply and in some cases they make no effort to disguise the fact that they’ve done this because you'll find different font sizes and styles. I don't know whether they think that people won't notice this because they certainly do. But at the other end, and it's sometimes often the case that it's maybe the parties and the individuals who have got little chance of electoral success that have been, have taken the most effort to respond more fully. Maybe because they are just that desperate for a vote, any vote.

**I: So that's actually successful then because that's what you were looking for, I take it, the actual engagement, 'this is what I think, that's a good question' or 'this is the focus of what we're trying to achieve with our policy'.**

M: And also whether there are any efforts for the conversation to continue. I certainly do – Ukip, bizarrely, were one of the more engaging ones and one guy, he sent me his mobile number and offered to ' phone me up if you want to chat further' and one guy I remember sending us a video clip that he'd just recorded in his study at home, two or three minutes long, giving his personal thoughts. So a lot of shining examples like that but on the whole you are either ignored or it's a standard response. And I think over the years maybe the Green Party have probably been the best at responding and making efforts to make it – add that personal touch.

**I: Rather than just a standard supply?**

M: Yeah.

**I: Again, I think we've covered this: your current research questions and again we've gone through the next question as well which is the role social media plays in your current research. So again it's really monitoring, well not monitoring but studying the engagement for various topics, whether it's with the people who are using social media to give out information in terms of politics or whether there are – the public is engaging with what they put out in terms of politics. So I'll move on to the kind of tools that you may be using. What tools do you currently use to capture or extract social media data?**

M: It's always been a problem as far as I'm concerned. In terms of Facebook it always has been (and blogs as well) and remains the case that I just have to capture, copy and paste.

**I: Like a screen capture or…?**

M: I tend to…yeah, you know when you get Facebook you'll see there's maybe thirty-five replies, open up the entire list and then just copy and paste, drag the mouse over the text and copy and paste it into a Word document. Now I'd love to know if there's a tool out there somewhere that can do this, I haven't encountered one so far but that's always been the case, just the manual capture.

**I: But then I imagine this is very focused then because you can identify – for example, say a politician posts something about a news story or about a policy or a debate in the Commons, you can then see if there are a certain number of replies and as you said, go to the Facebook page, open it and expand all the comments. And that’s kind of like a good snapshot of how people are responding?**

M: Yeah, yeah. What we've tended to do is because it's largely been based on election campaigns, it's maybe to look at the four weeks leading up to an election day and identify which parties and candidates have Facebook sites and Twitter posts and just capture everything that has gone on on those sites within the four week period. Now the very first time I tried it, it was my own fault, I think I underestimated, even back then, how much traffic there would be on it and I said, 'ok, I'll do this on the eve of election day', so just on the final day, just do a retrospective look back of the last four weeks. But a combination of the amount of posts and the copy and paste approach, it was my first time at it and unknown to me I was capturing all of the background coding and hyper-links and so on, and that all added to the time. So you could highlight two hundred responses to a post, do the copy and paste and when it came to the paste the machine just whirred and whirred around for ages before it finally deposited it. And I ended up having to stay up right through the night of election night. I would have probably done that anyway to watch the election coverage but I was copying and posting and so having learned that lessons subsequently I had – I normally capture the four week period so I'd maybe capture the first three weeks and then just leave the last week for the day before election. That made it a bit more manageable. Also learnt don't copy and paste everything, just copy the text and plant that down so, ok. Again, that technique is working ok but last year it just went crazy and I'm not sure anyone could have estimated just how active things got on social media– you've to open it up, you can only open up 100 posts at a time and that just took an eternity.

**I: So when you are capturing all the data and you are copying and pasting it, do you then have to process your word documents afterwards as in tidy up the text, make sure everything is in a manner so it's organised. Does that happen automatically with the copying and pasting?**

M: I guess the coping and pasting is done chronologically, so it would be chronological order.

**I: But you don’t have any, say, indentations which you want to fix?**

M: No, I think with even just the text-only one, the initial post will appear far left and then the subsequent posts will be indented and responses to those as well. So you have a…

**I: So it is fairly organised then?**

M: Yes, it should be.

**I: There's not much post-processing afterwards, once you've captured it then you can move onto the analysis, that's what I was trying to find out; whether there was any further post processing that you have to do?**

M: I'll guess we'll move onto this in a wee while, in terms of the nature of the analysis, just the approach that I use. I really just require the text itself, I just want to know what is being said, yeah, that's the main thing. So that's Facebook, so that is still a largely manual and time-consuming task. Interestingly as well I did find that during the referendum as well that, it seems to be all to do with Facebook's algorithms and so on that even though you asked for all posts to be displayed, it didn't. I would get additional ones that had been posted weeks earlier that hadn't…appeared with my initial opening up of that. So to try and be as thorough as possible…

**I Is this via Facebook itself?**

M: Yes, uh-huh.

**I: Did you look into how people are using, say, Facebook and they are linking it back to an external website or linking comments from an external website to Facebook? Would that perhaps be…?**

M: That forms part of our analysis because we…I guess – will I talk about the analysis at this stage?

**I: Yes, that's fine, that's fine.**

M: I guess the coding scheme I've devised over the last five, six years looks at two things; it looks at the nature of the communication taking place, whether it's just the one-way broadcast of information from the politicians to…

**I: Or whether it's a dialogue?**

M: That's right. And also subsequent analysis looks at the subject matter being discussed and also I break that down into whether, it's what I call primary broadcast posts; this is where the politicians…

**I: (…)**

M: Yeah, there we go. And then alternatively their secondary broadcasts posts where they are simply linking to a news story or website or something like that.

**I: So primary and secondary broadcasting is the way that they are…and you used this to describe the type of content that it is?**

M: Yes, I guess that is more the nature…we'll call it primary broadcast, secondary broadcast, I have a dialogue and engagement category which is where there is clear, two-way dialogue going on between politician and follower, or followers and then I have unreciprocated engagement but that largely just relates to Twitter traffic and this is where the politicians are following – and it's often celebrities, or something like that and what they call commentators. They'll have seen a post on the celebrity's Twitter post and they've then tried to engage with then hoping that a discussion starts. It rarely does, hence the term **'unreciprocated engagement' because they tend to be ignored completely.**

**I: And do you check the kind of people who are targeted from this sort of response from the perspective of the politician?**

M: I guess not to that detail. We haven’t really explored, let's say, what are the political opinions of the celebrities and looked for any relationships between them and the party of the individuals we've tried to contact, I haven't gone to that extent, it's largely just been a quantitative….

**I: To study the activity, I suppose?**

M Yes and to see what has gone on. That has dropped, it has to be said, in the most recent study; that they seem less inclined to waste their time in trying to contact people who are likely to ignore them. So there's that, there's the communication, nature of the communication, those are the four categories and there's the subject content of these posts which are broken down – it's a broad subject breakdown, things like are they talking about their own personal campaign, activities, are they talking about what is going on nationally, are they talking about news coverage of the election or are they talking about other current affair stories. There's a broad subject coding scheme that I use as well. So it's looked at in two ways.

**I: That's very interesting, I showed you this honeycomb framework so whilst we're on the subject of analysis, so you think this is something that would be useful in guiding your capturing of the data? Because what we're currently looking at is developing some queries for accessing Facebook and the tool that I've developed at the moment, what it does is it sends off a URL to the Facebook (API) and then that returns, say, the posts, the likes, the links and what we've discovered is you can just attach onto this query string every possible field you have for public profile on Facebook but the thing which we're concerned about is the information overload, so to try and constrain it we were considering using a framework such as this or the coding scheme you were talking about, that seems very, very appropriate. But I was wondering whether you had any thoughts on this framework and if we had queries which were built on, say, we could have frequency of activity within these categories, so frequency of primary broadcasts, frequency of secondary broadcasts – would that be something…?**

M: Yes, I guess so.

**I: Or do you think it's a bit too high level or…a little bit vague?**

M: No, in terms of high level, perhaps, although having said that, if this were done automatically this would save considerable time in manually – I should point out, I'm coding the post, let's say, in a single Facebook page then what I do is use the old five-bar gate method just with paper and pencil, just manually tallying up what post gets allocated to which code.

**I: And when you've got the frequency per codes, I imagine…**

M: I then input that into SPSS for further analysis. When I say 'further analysis', just producing a nice bar-chart or something like that.

I: So that's interesting because what we're moving towards is, at the point of **capture, once the data has been captured, we then present the user with a knowledge-graph which will show you, say, the distribution of Tweets for, like a particular hashtag, a particular public profile or the distribution of re-Tweets or just some general descriptive statistics about the sample so that the idea is the social scientist understands 'this is the data I have, these are the high-level assumptions we can take to do further analysis', does that sound something…?**

M: I think if it were Facebook I would probably find that more useful and the reason I say that is with Twitter, until maybe two, three years or so ago I was using the same approach, just the copy and paste and because I could not find a tool out there that seemed appropriate for my needs. But over the last two or three years I've started using a package called 'Twitonomy', I don’t know if you've heard of Twitonomy?

**I: Oh yes, yeah.**

M: It will go ahead and capture all of the Tweets within a particular Twitter account for me and it comes up with basic statistics like that and although of course it captures all of the data in a very handy Excel form, in terms of the coding and analysis that I do, I still have to go through each Tweet manually to code it.

**I: And do you use any software to assist you with that or do you have the Word documents and Excel files and you are keeping a manual note elsewhere?**

M: That's what I'm doing, yeah. I'm still unconvinced about the usefulness of any tools out there that might claim to…

**I: To automate this for you?**

M: Yeah. And I've seen, including papers in the field that I'm in, the political engagement, there's been a few recently that have undergone the sentiment analysis-type approach.

**I: You are a bit sceptical about that…?**

M: I am, yeah, I think just because I feel it doesn't pick up on sarcasm and irony and so on.

**I: To interpret it?**

M: Yeah, especially of the geographical elements to know, ok, are they talking about an area that is within their constituency so I still felt that human effort was necessary. We focused on political engagement for our research and there was a paper came out maybe three years or so ago looking at MPs' use of Twitter and they had used an automated approach but it seemed largely based on how many times, how many re-tweets they'd made and let's say for levels of engagement there was using the @ symbol as evidence of them engaging with others. And I just felt that was a fairly basic and crude measurement.

**I: There's too many assumptions behind the use of (*that),* would you agree with that?**

M: I think so, yeah. And without looking at exactly what was going on within the rest of the – because I think what I have found as well is in going through the manual analysis, when politicians do enter into a dialogue with users, it does tend to be more often with people they already know, either personal friends or party activists and so on, rather than a complete stranger. And this is borne out with our sending enquiries by Twitter because they largely get ignored because the politicians say, 'ok, who is this person, no idea who they are?' And they'll refuse to respond. I think just using the 'at' symbol as a measure of engagement and dialogue doesn't really give the full picture.

**I: So what we're also thinking of is at the point where you have a load of Tweets or a load of posts and you've done your frequency analysis, the more quantitative stuff, now whether that's provided through a knowledge graph or, say, as it's done with Twitonomy, presumably you would move on to coding the actual content and looking at themes. So what we're interested in is how you would code the themes and, say, after coding 10% of the sample if you could click a button and the machine says or the software says 'based on your current coding sequence I'm x percent short but I can code the rest of the data of you, as you've done', and at the end of that process, depending on the accuracy that the machine can predict your coding it can then code the rest of the dataset based on the codes you've already started off with, as in the manual thematic codes, would that be something you'd be interested in or would you be a bit sceptical?**

M: I think I probably would be sceptical, it would be brilliant if that were possible.

**I: So we've also, haven spoken to other social scientists and people working in this area, we've anticipated that it's a little bit of a leap of faith because it's almost like a black box can somehow reproduce what I've been carefully going through in terms of my knowledge of the subject area. Now what we're also interested in doing is having to restore a degree of faith, having the machine identify particularly troublesome cases which it finds difficult to classify. And there are some algorithms out there which will be able to detect sarcasm or irony but just flag it, so again, would that be something that you would consider using or would you still be a bit more…?**

M: I certainly wouldn't dismiss the idea completely, I mean I think I would be interested in seeing the end results of the analysis and maybe just blow away all my scepticism, absolutely.

**I: But one thing we're also very clear about is the – well not very clear about actually, is how your workflow would actually fit into a tool because we're focusing on capture at the moment so you mentioned that you identified the topics, the politicians, the individuals with accounts on social media who you are interested in and then you began by just trying to capture everything all at one but then you moved to a more manageable approach where you split it over time. And once you have the data in Word documents or an Excel file then you still go into this manual coding process and obviously this is very time consuming. So what we're trying to do is deal with the trade off in terms of increasing the sample size to the point where manual coding, it's still an option but you'll only be able to look at so much of the sample and then training a classifier to allow you to apply that coding sequence to a degree of accuracy to a much larger dataset. So would you be convinced with that kind of trade-off?**

M: I think potentially, yes. The more recent research I did, , again I stuck to a four week data collection period, I guess to compare it with the periods that we've collected during the election campaigns but I'm more than conscious that all that we need is one big news story to come out during that four weeks that could skew the entire thing, completely. So in many respects it would be good to have a more longitudinal approach looking at maybe a year's worth of Twitter so that you get a better and…

**I: And that would be good motivation than it would be interesting then to see how far we could push these automated classification techniques and obviously with the condition that the machine actually identifies, or the software identifies, particularly difficult cases which do require manual analysis.**

**So there's a couple of other things; you mentioned how the honeycomb framework could be potentially useful?**

M: Yes, there is other aspects of that, certainly the conversations, it's very useful, I think particularly between the Twitter account holder and their followers. Less interested in what conversations go on between individual followers of a politician. There's certainly a lot of that goes on on Facebook, I'm conscious of that, I do – well in past years before the referendum where the traffic was just too big I have attempted to quantify that, how much conversation is going on between followers, that the politicians themselves were not involved in.

**I: So would you be able to say – classify these conversations using..could you fit your existing coding sequence into this framework? So here with conversations, could you have 'these were derived from primary broadcasts and these ones are derived from secondary broadcasts' and then maybe do a correlation analysis between the frequency of particular topics?**

M: That would be really interesting if that were possible in an automated fashion; to see which subject topics generated the most discussion amongst the politician's followers, yes.

**I: And there was something I wanted to ask you about linking the data. You mentioned that you look at blogs as well, so do you ever consider linking data which you've captured from social media to other sources? So, say for example, as you mentioned you could be studying a campaign and then something will come out in the press and that will suddenly increase the number of responses or also potentially reduce the number of responses or focus on a particular politician or political party?**

M: I think I'm kind of conscious of that as I'm going through the analysis, that there may well have been a particular story or episode that generated a lot more traffic than others. And in many cases it would probably just be a mental note to self along with just a little handwritten note on the coding sheet saying 'ok, this particular issue caused a lot of comment.'

**I: So there could have been a sudden increase in frequency of posts and re-Tweets because Question Time has been on, a leaders' debate or something?**

M: Yes, I guess that's one thing that we have done, I guess during the election campaigns is looked at when the period we've covered has coincided with the leader's debates, I've taken a separate note of how many Tweets on the politicians' websites were made during or immediately after these events. afterwards and so it was clear that the more traditional media of television was driving a lot of what was going on on Twitter.

**I: Would you say that is kind of like linking the dots together because you are looking at how a certain event is actually, as you say, increasing the traffic. And perhaps if you – I'll suggest, would it be plausible or interesting for you to look at a certain hashtag and then see how the frequency of a certain term or hashtag related to policy, how that increases in terms of re-Tweets, people posting about it when prompted through these external events and other media?**

M: Yes, I guess to a certain extent we were doing that with the research that I mentioned, , we were looking at what was going on and for that we had the assistance of our colleagues in computing

**I: Yes,.**

M: Yes, the package that they had. And we gave them the hashtags that the programmes themselves were promoting in advance and we just focused on the use of those but that might be an interesting approach as well; to just focus on a hashtag perhaps or a subject area that is not limited to a particular event such as a leader's debate and trace the flow and the use of that word or that hashtag throughout an extended period to see if certain events prompt it to rise or fall.

**I: In terms of your workflow, do you ever, at the point where you've captured, say, a load of Facebook posts, would you then look at how the same profile on Twitter, would you ever compare posts for the same person or would that be a bit too…?**

M: What I **do** do, it's on the coding sheets, we have a separate category and again, it's to prevent duplication and so this account is for Facebook sites that are potentially just feeds from Twitter accounts so it's clearly the same content…

**I: So you have to flag it for duplication reasons?**

M: Yes, so that's done, again that's manual as well and similarly if the Tweet is simply a link and the first few words of a Facebook post, it's noted as such. I think the category I use has just got links and feeds to and from other social media sites, so that's quantified so that the content of one of them will be analysed and note taken that…it's duplicated in another social media application. But again, if that were possible – more easily and quickly using an automated method, that would be most advantageous as well.

**I: Ok, I've just got a few other round-up questions to ask. Are you aware of any software at the moment which could actually assist you with capturing and analysing social media data, commercial packages like NVivo, have you?**

M: Well I'm aware of it and nCapture as well, that's the add-on for NVivo?

**I: Yeah.**

M: I haven't really seen that demonstrated at all, I know one of my colleagues looked at it briefly prior to us doing the general election, I can't remember what his conclusions were but I think the feeling was that it didn't really match what we were looking to do.

**I: As in it would be difficult to use this tool within the framework of your layout and coding sequence?**

M Yes, I think, if I remember correctly, I think also as well – and this is related to the Twitter traffic during the TV debates – that a lot of what we were looking for was very quick results, quick, cheap, dirty results that we could rush out on a press release that same night and this is what a lot of the press releases were based on – the Twitter traffic went through the roof when Nigel Farage started talking about health tourism and so on. And so it was this very cheap and nasty quantitative approach that we were using the package for. And certainly the NVivo nCapture, I don't think that helped us with that at all. I think certainly when it came to the more detailed qualitative – yes, we're looking the content analysis aspect of if then, yeah, I think we could see that that could potentially be of use there. But it didn't go any further than just that initial exploration. Our friends at Computing, with their social sensor tool, it was kind of sold to us with the suggestion that it had a kind of content analysis aspect to it and that it would identify key themes coming out but I think in reality it was identifying which was the most re-Tweeted Tweet.

**I: So it was very – it was determined by frequency and presumably calculating 'this is very popular and this is being…'and it's counting the meta-data.**

M: I believe there is a public – it's in the public sphere, the tool itself as an example, and they had a kind of rolling symbol that came up with the top Tweets at any given time which is all very well and it looks very nice but not really what we wanted.

**I: Ok. Well I think it's very interesting the discussion we've had on capturing an analysis but I'm quite sure we've gone over this already but just to be clear: in terms of storing the data you capture, do you have any issues in terms of terms and conditions with Twitter, obfuscating, perhaps, user ids, anoymising things like that, do you have any issues with storing your social media data?**

M: This is a very interesting question and I would have to admit that until we got involved with Computing I was kind of blissfully unaware or maybe preferred to ignore any potential difficulties in capturing and storing Facebook and Twitter traffic, my assumption was it's in the public domain. And I think when I've captured and stored the data it still has all of the Twitter ids and names and Facebook book profiles things attached and it's on a hopefully secure university network. In terms of presenting the results, you tended to…

**I: You'd anonymise it.**

M: You anonymised everything. Except, let's say, more recently if there were particular politicians or parties that – I think it's more likely to be exemplars of good practice, I would name them in, rather than be openly critical. I have certainly included example posts that maybe don't paint particular parties and individuals in a good light but they've been anonymised. I've been cautious as far as that is concerned. But in terms of the storage of the raw data, no. And I know that our friends in Computing started going on about Twitter's terms and conditions and how they only allowed a certain percentage of Tweets to be captured before..

**I: Before they charge you.**

M: Ah, yeah. And it probably hasn’t been a problem with the manual, the traffic has been such that it probably didn't go anywhere near Twitter's and I'm assuming, maybe incorrectly, I don't know, that since I've switched to using Twitonomy to capture that if there were any problems then Twitonomy wouldn’t be able to capture that. But I'm also conscious as well that certainly incrementally over the years the social media traffic has grown and continues to grow and no doubt will continue to grow in the future but…

**I: So would you say, based on what you were just saying just now, that you can see this research expanding and ongoing because of the growing popularity and engagement?**

M: Yes, I think it may well be that we'll move on to different applications, that Twitter and Facebook will fade into obscurity and be replaced by others. The traffic, it's certainly continued to grow, again there may be a time when it reaches a plateau but I think from a personal point of view I think there has to be a point when I could only go along my current approach, which has been a kind of census approach; just capturing everything within that period, I think there would have to be a limit. I would be doing some stuff during next year's Scottish Parliament elections and I'm conscious that if the traffic continues to increase as it has done, I may have to introduce a sampling approach..

**I: And use it as a lens then? This is like – just an idea of the current comments at** **this one time?**

M: That's it.

**I: If the traffic became exponentially too big?**

M: Perhaps, instead of focusing on two weeks maybe just cut it down to one or instead of looking at all of the candidates across all of the parties maybe taking a selection from each party so just making it a more manageable number, somehow.

END