UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Electronics and Computer Science

Revealing the Content of the Edu-Blogosphere: Taking a Seat in the Virtual Staffroom

by

Sarah Hewitt B.A.(Hons), MSc

Thesis for the degree of Doctor of Philosophy

April 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
Electronics and Computer Science

Doctor of Philosophy

REVEALING THE CONTENT OF THE EDU-BLOGOSPHERE: TAKING A SEAT IN THE VIRTUAL STAFFROOM

by Sarah Hewitt B.A.(Hons), MSc

Teaching can be an isolating profession, but the growth of the world wide web has given some teachers and other Edu-professionals the opportunity to have discussions using social media, and to publish their own blogs online. Some of them have been doing this for quite a long time, resulting in a considerable corpus on the Web. Not only is much being said, but there is also evidence that some voices are being heard by people that have the power to modify Education policy in England.

This does present challenges for the researcher. In order to understand the content of the corpus, the posts must be harvested and subject to a methodology that will make some sense of variety of topics discussed. While some research has been carried out, a comprehensive survey of blog posts from 'independent' platforms such as *Wordpress* has not. Reading and coding by hand is no longer feasible with such a large data set; an interdisciplinary approach characteristic of Web Science which combines tools from machine learning in Computer Science with methods from Social Science is needed. Such a methodology has been used successfully to present an overview of the topics discussed in the Edu-blogosphere.

As well as researching what the Edu-community has been discussing over time, this research also shows *how* this has changed, and that there is a link between the topics of discussion within the Edu-community and changes in Education policy. It reveals a rich source of professional discourse that has much to offer members of the teaching profession, whatever their capacity; and indeed anyone seeking to understand the concerns of its members and to celebrate their success.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Sarah Hewitt B.A.(Hons), MSc , declare that the thesis entitled *Revealing the Content of the Edu-Blogosphere: Taking a Seat in the Virtual Staffroom* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: No citations

Signed:...

Date:..

Acknowledgements

So many people to thank, and I'll begin with the most important ones, my mum and dad. I'm sad that neither of them are here to take pride in the things I've achieved, because without whatever it was they gave me - genes, character, upbringing - I wouldn't be here writing this. They were very ordinary, working class folk who came from very poor households, and were, like so many of their generation, adamant that I would have all the advantages they didn't, starting with a good Education. Thank you.

Next, a big thank you to my supervisors, one of whom interviewed me as a potential PhD candidate right at the beginning. You've been there when I needed you with advice and encouragement. There are some other people (who shall remain anonymous) that were responsible for my application to do a PhD in the first place: some people know the story, I'm not going to repeat it here, but all I can say is if life gives you lemons, have a cry, go to bed, get up and make the very *best* lemonade you can.

Of course then there are my friends old and new: Hazel, Carolyn, Jackie; Nikko, Briony, and Nick. There are others. You know who you are. I've been very lucky to have been part of such a magnificent cohort. Thank you, one and all.

I must also mention Andrew Old, a blogger, tweeter and one of the central figures of Edu-twitter whether some people like it or not. He provided the list of URLs I used in this research, and saved me hours of work. Thank you.

Finally, perhaps the most heartfelt thank you must go to my partner, Christopher. He's been with me from the very start of this journey, and without his gentle presence and support the road might have been rather different. Millie (my dog) also thanks you because she's only had to go to kennels twice in five years.

Chapter 1

Introduction

1.1 Researching the Social Layer of the Web

It is easy to forget that social media is a new phenomena; it feels as it it has always been with us even though this has only been the case for people born since 1990s. 'Always on' broadband has only been widely available to most UK households since around 2010, a mere 9 years. Wordpress was launched in 2003, Facebook and Twitter in 2006. Since then, a whole raft of other social media platforms have arrived, although Facebook remains the most popular in the UK[1]. According to the London School of Economics, in 2017 39% of the population of the UK were 'active social media users'[2]. The importance of social media as a method of influencing voter behaviour was made apparent in the wake of Carole Cadwalladr's investigation into Cambridge Analytica[3]. While this demonstrated that social media users could be influenced, there are also plenty of examples where people have used social media to complain or express their opinions and make their voices heard such as the #metoo campaign on Twitter, which successfully raised the public's awareness of sexual harassment and sexual assault[4].

Underneath the headlines, though, communities have been forming, facilitated by social media, including teachers, lecturers and other Edu-professionals. While we can assume that not not every teacher in the UK engages with social media in a professional context, there is now a sizeable community that have been writing blogs and promoting them (usually via Twitter) for several years. There is now good evidence that, as well as discussing their profession among themselves, this social layer of the Web is also being

[1] https://www.statista.com/statistics/280295/market-share-held-by-the-leading-social-networks-in-the-united-kingdom-uk/

[2] https://info.lse.ac.uk/staff/divisions/communications-division/digital-communications-team/assets/documents/guides/A-Guide-To-Social-Media-Platforms-and-Demographics.pdf

[3] https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy?language=en

[4] https://www.theguardian.com/world/2019/mar/27/mexico-metoo-workplace-abuse-sexual-harassment-media

used to mount a critical response to reforms instigated by the Department for Education, which regulates Education in the UK. For the first time, the impact of these reforms can be now observed and measured at scale, which is the focus of this research.

1.2 Original Contribution and Research Questions

People - teachers - blog because they have something to say. It might be that they have a view to express, something to share, or because they feel the need to build an identity and a presence on the Web. As a group of professionals, busy preparing the current generation of young people to take their place in society, their voices deserve to be heard above and beyond the communities they build together. There is now evidence that this is the case, although how much actual influence they have is uncertain. What is certain is that there are many Edu-bloggers out in cyberspace, writing about all things that concern their profession, and in some cases they've been writing for over 10 years. Theirs is at the very least a rich source of shared ideas through communities of practice as well as expressions of concern and criticism of the Education system. This research will explore the content of their blogs at scale, bringing fresh insight.

1.2.1 The Role of Web Science

This research is undertaken within the relatively new discipline of Web Science. This encompasses "...computer science and engineering, the physical and mathematical sciences, social science and policymaking..." (Berners-Lee et al., 2006, p.4) and is described by Berners-Lee et al. as 'inherently interdisciplinary'. The web itself encompasses the following areas (Phethean et al., 2016):

- the *technical* layer: the technologies behind the web;
- the *social* layer: the people, the content they share, the social networks they create; and
- the *market forces and policy* layer: "...which relates to the interconnected economic and political factors that shape the Web's evolution".

The source data for this research is located in the social layer, traditionally the area of most interest to the Social Sciences. The research questions (detailed below) are typical examples of those asked by Social Science researchers. However, as will be explained, the scale of the data that will be used to answer the questions demands tools from Computer Science in general, and machine learning in particular. This combination also forms part of the unique contribution of this research, as a thorough understanding of several aspects of Computer Science is required in order to harvest, clean, process and analyse the data. Web Science allows for the combination of approaches from different schools to be brought to bear on research questions that do not ‘fit’ neatly into one area. There are no prescribed methodologies or approaches - the researcher is free to select methods that best suit the research question or questions. This pragmatic approach is discussed in more detail in Chapter 3.

The first question that will be addressed by this research is:

1.2.1.1 Research Question 1: What kind of subject matter, themes or issues have been discussed by the Edu-community since the arrival of the first available Edu-blog?

This is the main question that has been asked of the entire corpus. Two approaches to answering this question have been employed: the first draws on research that has already been carried out, using this to construct a framework of topics to map to the Edu-blogosphere; the second makes no *a priori* assumptions and seeks to extract an unknown number of topics. The research pertaining to both approaches are reviewed in Chapter 2. Chapter 3 discusses and details the chosen methods, and results are discussed in the Chapter 4.

It is practically impossible to pinpoint the very first Edu-blog. It is possible that the URL is unknown or has since been deleted. What is possible is to gather blogs from the *known* URLs, find the earliest, and proceed from there. Fortunately, a long-standing member of the Edu-community has maintained a spreadsheet of blog URLs[5] which catalogues over 2000 entries. Some newer blog URLs that do not appear on the sheet have been added to the list. Not all blog URLs were harvested, which is discussed in more detail in Chapter 3.

1.2.1.2 Research Question 2: Has the subject matter, themes or issues discussed by the Edu-Community changed since the arrival of the first available Edu-blog?

Grouping the blogs by year, and applying the framework across each year, can reveal how the content of blogs has changed over time. It may be that new themes etc. emerge, or that the same issues arise but take prominence over others from time to time. Once this becomes clearer, it will then be possible to triangulate this with a timeline of events in Education (policies, announcements, key speeches etc.) to see of there is any relationship, which is the focus of the final research question.

[5] https://teachingbattleground.wordpress.com/2015/08/12/please-help-with-the-uk-education-blogs-spreadsheet-version-12/

1.2.1.3 Research Question 3: Is there any link between the subjects, themes or issues discussed by the Edu-community and changes in Education policy?

While there has been some research into teachers' use of social media and blogs, this is the first time a large survey of the Edu-blogosphere, going back as far as the first available blog, and forward to the end of 2017, has been undertaken. A timeline will be constructed, and the results from research question 2 triangulated with this, providing a year-by-year visual representation of the themes etc. discussed by the Edu-community.

1.3 Why Research Edu-Bloggers?

I qualified as a secondary school teacher in 2004, at around the same time that Information Technology (IT) had become a focus in schools. Teachers, at least in the UK, were being issued with a laptop computer, and electronic interactive whiteboards were being installed in secondary schools.

Like most other teachers, my experiences were mixed. Teaching is always challenging, but as more and more emphasis was placed on good SATs[6], GCSE[7] and GCE[8] results, the pressure and workload increased, often alongside class sizes. Behaviour in the school I worked at was often poor, especially with the kind of disruption teachers refer to as 'low level' such as calling out. There was pressure from local education authorities on schools *not* to exclude the most disruptive pupils, all of which meant that I and my colleagues worked incredibly hard, were frequently exhausted, and generally stressed. We were not alone (Kidd, 2013; Greene, 2013a, 2016).

For various reasons, I also found myself questioning the teaching methods I'd been taught during ITT[9]. One of the places I looked to for support and advice was the burgeoning community of teachers on Twitter. Through them, I discovered blogs written by teachers and other Edu-professionals that demonstrated that I wasn't the only teacher working in a 'difficult' school, and questioning the orthodoxy[10]. In fact, as I will explain below, it turns out that I wasn't the only one reading what some high-profile members of the Edu-community had to say about the growing problems in Education.

Around the time I left teaching, Ofsted[11], under the Secretary of State for Education Michael Gove (Secretary from 2010 to 2014) was accused of unfairly criticising schools,

[6] Standard assessment tasks given at the end of KS2 and, up until 2008, also at the end of KS3.

[7] General certificate of secondary education

[8] General certificate of education (usually referred to A-Level)

[9] Initial Teacher Training.

[10] https://www.theguardian.com/education/2009/jan/05/ofsted-boring-teachers?

[11] The Office For Standards In Education, Children's Services and Skills, a non-ministerial department of the UK government reporting to Parliament via the Department for Education.

and returning unfavourable judgements. The inspectors were looking to observe specific teaching methods, often referred to as 'child centered' (Peal, 2014). At the same time, the government were encouraging failing schools to become academies[12]. Some members of the Edu-community began to criticise Ofsted through their blogs (Peal, 2015, Introduction) and there is some evidence to suggest that these blogs were read by government ministers in the Department for Education [13]. In any event, Ofsted were instructed to change their inspection criteria[14],[15]. It appears as if the voices of those delivering education day-to-day in schools were heard at last. The Edu-blogosphere was also mentioned recently by the current Minister of State for School Standards, Nick Gibb [16]. Sean Harford, Ofsted's National Director of Education is active on Twitter and frequently engages in discussions about Ofsted.

Blogs reflecting 'teacher voice' have become an easily accessible way for politicians to see what is concerning the Edu-community, as well as a useful way for teachers to engage in debate and share ideas. An initial sweep of the existing research into teachers' blogs revealed that some work had been done looking at why teachers blog and what they blog about, but on a small scale and certainly nothing to address blogging over time. This research will address these issues by presenting an overview of the topics discussed by many hundreds of bloggers, starting from 2004 (the date of the first known and accessible blog) to the end of 2017. The results will also be triangulated with a timeline of the major events from Education e.g. changes in policy, key speeches by ministers etc. Advances in computer technology and machine learning have made it possible for such a large set of data to be processed using computer algorithms, techniques that have only become available to researchers working on non-specialised computer hardware relatively recently. A broad and far-ranging overview of blogs from the Edu-community is now possible and will provide an insight into the concerns, challenges and successes of the Edu-community over more than ten years.

[12] Academy schools are no longer under the control of the local education authority, and receive their funding directly from the government.

[13] https://www.theguardian.com/education/2013/dec/03/michael-gove-education-dominic-cummings-policies-oxbridge

[14] https://www.tes.com/news/school-news/ofsted-watch/ofsted-scraps-grades-individual-lessons

[15] https://policyexchange.org.uk/publication/watching-the-watchmen-the-future-of-school-inspections-in-england/

[16] https://www.gov.uk/government/speeches/nick-gibb-englands-education-reforms

1.4 The Rise of Social Media and the Benefits for the Edu-Community

Social media can be defined as "...a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (Kaplan and Haenlein, 2010). Through virtual communities, people can share ideas, opinions, resources; create artifacts such as videos or images to share; and contribute collectively to any or all of these things. Social networking sites facilitate the creation of networks by allowing users to create links[17].

As well as the creation of communities, virtual spaces also allow people to generate *content*. One of the first blog-publishing services, started in 1999, was Blogger (now owned by Google). The blogging platform Wordpress came along in 2003[18]. Facebook was opened up to everyone with an email address in 2006[19], with Twitter launching earlier that same year[20]. Although Facebook and Twitter remain the two most popular social networking sites, many others have since been developed e.g. YouTube, Instagram, WhatsApp, and Weibo in China.

Blogs are social media Web postings listed in reverse chronological order. They are mainly textual, but increasingly they now include images, videos (home-produced or embedded links from sites like YouTube), or infographics (digital posters). The term *blogosphere* was first used by Brad Graham in 1995[21] (original reference now unavailable) to describe the network of blogs that were beginning to form. In 2010, Technorati estimated that there were more than 8 million blogs online, with a new blog created every 7.4 seconds[22].

As Rebecca Blood wrote in her article, when she started blogging, bloggers created their own Web pages, and entries "...were short, usually containing links to the larger Web and appearing together on one long page" (Blood, 2004, p.54) . She recalls that bloggers "worked hard to become dependable sources of links to reliably interesting material"; the message was shaped by the medium by incorporating links, writing concisely, and "rooting out interesting material" (ibid, p.55). The reverse - the medium shaping the message - began to happen from 1999 when Blood says "everything changed" with the arrival of Blogger.

[17] 'Links' are formally known as 'hyperlinks', which enable the reader to jump to another document on the www, or a specific point within a document. Thus, users are able to browse the Web

[18] https://en.wikipedia.org/wiki/WordPress

[19] https://en.wikipedia.org/wiki/Facebook

[20] https://en.wikipedia.org/wiki/Twitter

[21] https://en.wikipedia.org/wiki/Blogosphere

[22] http://technorati.com/state-of-the-blogosphere-2010/

Blogger made it easy for new would-be bloggers to set up their own page, but it wasn't until Web pages had permanent URL links[23] (permalinks) from 2000 that they began to add links to other URLs. This was quickly followed in 2001 by the ability to leave comments, and 'trackback' (sometimes called *linkback* or *pingback*) which allowed the blog to notify another blog of an update. Platforms like Wordpress make it very easy for anyone to set up their own blog, and to find others. Wordpress even encourages users to set up a 'blogroll' on the home page of their blog, so that visitors can see what interests the author, and perhaps choose to read one or two and link to them in turn.

Readers can usually leave a comment at the foot of a blog post (although the 'owner' of the blog may choose not to show comments, or only show comments when approved by them). The act of leaving a comment can include leaving the name of the commenter (although it is sometimes possible to leave comments anonymously), which again creates a link back to the commenter.

The content, quality and persistence of blogs and bloggers varies widely. For the social sciences, they represent a rich source of what interests humans, especially when bloggers begin to form links with each other and set up communities. Some of this will be discussed in more detail in Chapter 2.

1.4.1 Why write a blog?

There are many different reasons for blogging. Journalists often maintain a personal blog as a way of showcasing their work[24], especially as more and more of them now work on a freelance basis. People from other professions such as lawyers may use a blog to discuss certain aspects of the law[25]. Ordinary people have opinions they want to express, or sometimes just want to keep a kind of personal diary, even though that diary is in effect public. Teachers and other Edu-professionals maintain blogs for exactly the same reasons - they have things to say about their own practice, or policy, or have useful information they want to share; a blog is a good place to do it. This is expanded on in Chapter 2. Of course, not all teachers want to write blogs: some just want to read what others have to say, and may just choose to comment either directly under the blog post, or perhaps via Twitter.

Why people write blogs can be divided into two areas of interest: the *content* of the blogs (Schiano et al., 2004), and construction of an *identity* or *identities* behind the words (Schau, H.J. & Gilly, M.C., 2003). As this research is concerned with the content of blog posts from a specific community, the literature presented here is likewise focused on content as a source of data to explain why people write blogs, although there will be

[23] Uniform Resource Locator; or the www address of another point on the Web.
[24] https://www.facebook.com/pestonitv/posts/
[25] https://thesecretbarrister.com/

some discussion regarding identity and how it relates to specific categories of blog posts from the Edu-community.

A review into 'blogspace' was published in 2004 (Kumar et al., 2004). This focused on the Web site LiveJournal[26] which describes itself as "...a community publishing platform, willfully blurring the lines between blogging and social networking"[27]. This is a 'blog service' site as defined by Schmidt (Schmidt, 2007, p.1417), as opposed to sites like Wordpress which Schmidt describes as 'blog script packages' because the user has greater control over the appearance and data. While only researching one blog space, the study nevertheless was able to look at the entries of 25,000 blogs and revealed the "...rich and complex social environment..." (ibid, p.39) which existed. Although not expressly concerned with the content of blogs except as a way of discovering groups or clusters of bloggers, the variety of topics discussed was extensive (ibid, p.37). When the same study went on to look at the links across blog postings from individual sites hosted by, for example, Blogger, the authors found evidence of communities forming across platforms (ibid. p.38).

Another early paper 'Blogging by the rest of us' (Nardi et al., 2004) distinguishes between the 'heavy-hitting' blogs written by professional journalists, and blogs written by 'everyone else'. The study followed an ethnographic approach, interviewing 23 people. Most of them used their blogs as personal journals, and the topics discussed were wide. For some bloggers, the idea that their blog may not be read by anyone else wasn't an issue, as they were treating them more like diaries. Most, however, *were* aware that their blogs were read, which had an impact on how they wrote e.g. avoiding bad language if their blog was read by family. Many of them had already started to build a 'community' by finding, and linking to, other blogs of interest.

In a subsequent paper, Schiano et al. (2004) looked in more detail at why people blog. As well as the social aspect, the authors considered object-orientated activity i.e. blogs may be written with a specific object or purpose in mind, such as:

- to update others;
- express opinions or influence others;
- seek the opinions and feedback of others;
- to "think by writing";
- to release emotional tension.

Interestingly, one of the reasons given for blogging was the 'limited interactivity' offered via comments and replies, although feedback was often given (and welcomed) via other channels such as email or instant messaging (similar to Facebook chat today). In their

[26] www.livejournal.com
[27] http://www.livejournal.com/about/

first paper, the authors also identified bloggers who were using their blogs as a forum for "...ongoing work projects..." (Nardi et al., 2004, p. 1146). This suggests they were beginning to form 'communities of practice', which was mentioned above, and will be discussed in Chapter 2 in more detail.

1.4.2 Education Policy and 'Teacher Voice'

As a Guardian newspaper article made clear[28], the "...lack of a clear role for education professionals in policymaking" was an issue when the article was written in 2013, and arguably still is. The *Headteachers' Roundtable*[29], referred to in a later Guardian newspaper article[30] is evidence of the use some have made of social media to try and get teachers' voices heard. This is a group of Head teachers who have formed a thinktank[31] and write a blog collectively[32], but of course many of the blogs written by the teachers used in this research are trying to do the same thing - to express important, considered and *professional* points-of-view, and engage in professional activities such as Continuous Professional Development(CPD), in order to demonstrate directly and indirectly that they have views that are worth taking into account. Sam Freedman writes in his blog "But I still think it's exciting that for the first time ever any teacher anywhere can sit down and write something that could shift national policy."[33]. However, 'having some influence over policy' is not the same as having a formal role in the consultation process; and is not the same as a *practicing teacher* having that formal role. Even the Chatered College of Teaching, which was set up in 2017 and initially created to represent teachers, has seen some of the key positions go to non-teachers[34],[35].

The Edu-blogosphere is the area of interest for this research, and will be discussed in much more detail in Chapter 2. However, it is worth mentioning one of the research papers that got to the heart of why I wanted to research this area in more detail: 'Investigating Teacher Voice Through Blogs: policy, practice and local knowledge' by Greene (Greene, 2013a). Greene makes the argument that blogs written by practicing teachers (in her paper she focuses on K-12 teachers in the USA) are "... a way to "see" into the classroom" (ibid, p.1) and 'hear' the voices of teachers. These voices are largely absent from the process of devising Educational policy at both national and local levels. This early paper was developed further in 2016 (Greene, 2016). Greene makes the argument that American public schools are failing, citing high-stakes testing,

[28] https://www.theguardian.com/education/2013/dec/03/michael-gove-education-dominic-cummings-policies-oxbridge

[29] https://headteachersroundtable.wordpress.com/

[30] https://www.theguardian.com/education/2016/apr/26/headteachers-fightback-government-reforms

[31] http://htrt-thinktank.co.uk/about-us/

[32] https://headteachersroundtable.wordpress.com/about/

[33] http://samfreedman1.blogspot.com/2013/04/

[34] https://chartered.college/team

[35] https://teachingbattleground.wordpress.com/2018/10/06/the-chartered-college-of-teaching-a-broken-promise-to-teachers/

constant changes in the curriculum, and chronic under-funding. These are similar to the issues faced in UK schools [36], although from reading through some of the blog extracts in Greene's work, and the work of other Edu-researchers in the USA, my impression is that the situation in the UK isn't yet quite as bad. Certainly, the decision making and funding structures, with additional layers of bureaucracy at state level, make the US system more complex. The Web site for the US Department for Education states "Please note that in the U.S., the federal role in education is limited. Because of the Tenth Amendment, most education policy is decided at the state and local levels."[37]. This contrasts with the UK, where the Department for Education is responsible for education policy.

Greene sees blogs by teachers as filling the gap between policy and practice - as a way of informing the people responsible for making the decisions about what happens in the classroom of the *impact* of those decisions. There is now good evidence that this is happening in England (Wales and Scotland having greater autonomy), as evidenced by a speech made by Michael Gove in 2013, in which he stated "I'm a great fan of Andrew Old, whose brilliant blog Scenes from the Battleground provides one of the most insightful commentaries on the current and future curriculum that I've ever read..." and "I also hugely enjoy the always provocative work of Tom Bennett, the Behaviour Guru..."[38]. This, and other evidence is presented by Old in Peal's book *Changing Schools* (Peal, 2015, Ch.5).

A minister may, of course, only be interested in reading blogs that support his or her policy, looking for corroboration rather than challenge, but this is an issue outside the scope of this research. Here, the intention is to answer the research questions below, and reveal the scope and impact of the voices of the Edu-community over time.

1.5 Guide to the Following Chapters

In summary, there are many thousands of blog posts, going back over ten years, to be explored and analysed. The improvements in computer hardware, together with the development of algorithms and tools from Computer Science, means that it is now possible to do this. On the one hand, drawing on existing research into Edu-blogs means that we already know something about the topics discussed, and can then see how the proportion of each identified topic changes over time (and link them to events in Education using a timeline). On the other hand, we can use an algorithm to explore the blogs and let the topics emerge with no

[36] https://www.tes.com/news/even-tories-are-now-angry-about-education-funding-cuts; https://www.tes.com/news/are-you-teaching-zombie-lessons; https://www.tes.com/news/how-much-testing-too-much

[37] https://www2.ed.gov/policy/landing.jhtml?src=pn

[38] https://www.gov.uk/government/speeches/michael-gove-speech-to-teachers-and-headteachers-at-the-national-college-for-teaching-and-leadership

prior assumptions. Either way, this research will present a sweep of the Edu-blogger landscape based on what we already know and what new insights may emerge.

The following chapter, Chapter 2 is divided into two parts: Part 1 reviews the existing research into the Edu-blogosphere; this includes *communities of practice* and *personal learning networks* of which blogs are a constituent part. The *policy cycle* of Government is also included here to show where there is provision for teachers' voices to be heard. Part 2 presents research into content analysis from the disciplines of the Social Sciences and Computer Science. Figure 1.1 fulfils several functions with regard to the following chapters. Chapter 3 is divided into three parts, explaining the work undertaken to harvest, clean and transform the blog text data before content analysis can be done; content analysis; and the construction of the timeline and data visualisations. The final chapters discuss the results, a more detailed discussion of the results in Chapter 5, followed by *Conclusions and Future Work* in Chapter 6.

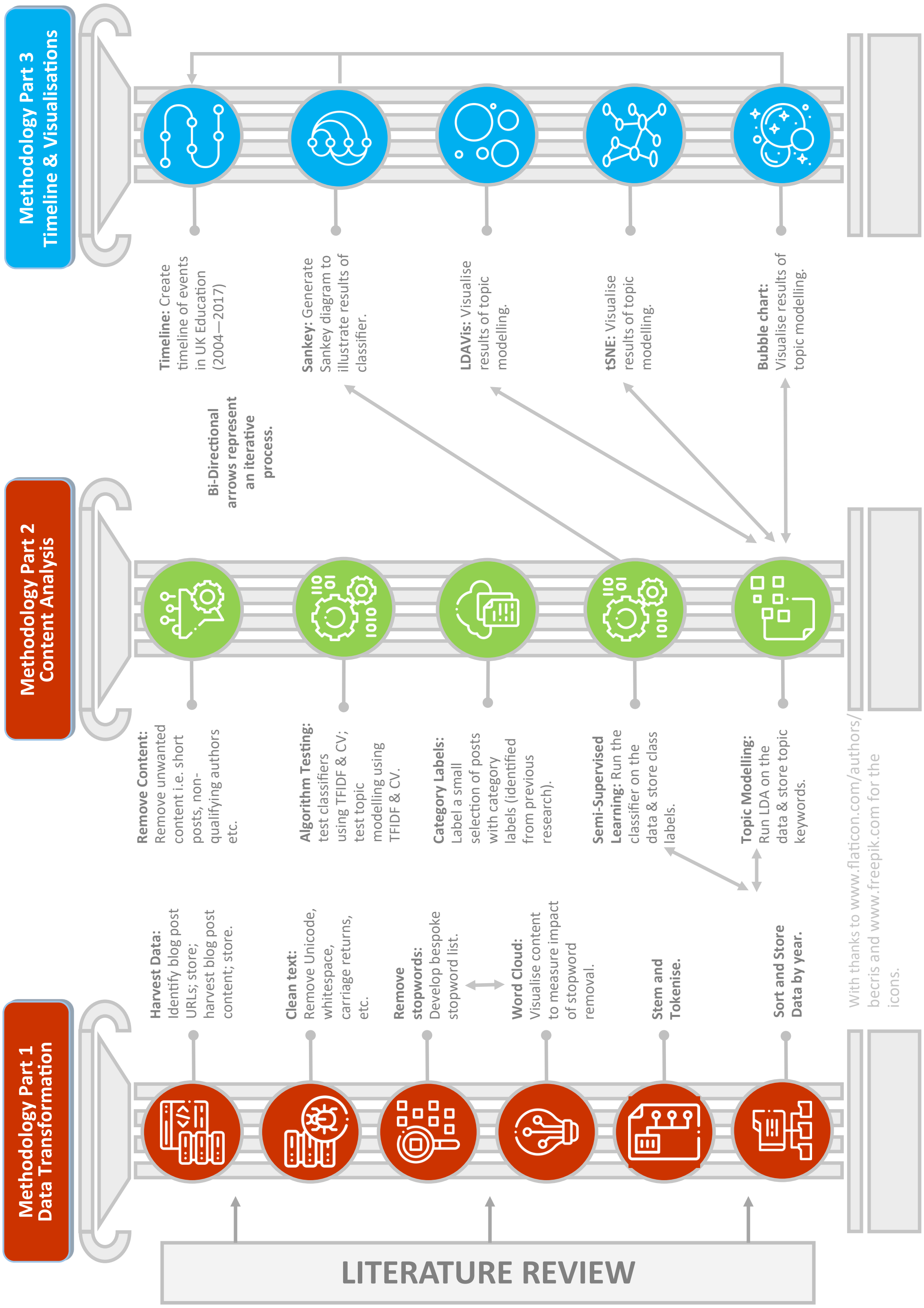

Methodology Part 1
Data Transformation
Harvest Data: Identify blog post URLs; store; harvest blog post content; store.
Clean text: Remove Unicode, whitespace, carriage returns, etc.
Remove stopwords: Develop bespoke stopword list.
Word Cloud: Visualise content to measure impact of stopword removal.
Stem and Tokenise.
Sort and Store Data by year.
LITERATURE REVIEW
Methodology Part 2
Content Analysis
Remove Content: Remove unwanted content i.e. short posts, non-qualifying authors etc.
Algorithm Testing: test classifiers using TFIDF & CV; test topic modelling using TFIDF & CV.
Category Labels: Label a small selection of posts with category labels (identified from previous research).
Semi-Supervised Learning: Run the classifier on the data & store class labels.
Topic Modelling: Run LDA on the data & store topic keywords.
Bi-Directional arrows represent an iterative process.
Methodology Part 3
Timeline & Visualisations
Timeline: Create timeline of events in UK Education (2004—2017)
Sankey: Generate Sankey diagram to illustrate results of classifier.
LDAVis: Visualise results of topic modelling.
tSNE: Visualise results of topic modelling.
Bubble chart: Visualise results of topic modelling.
With thanks to www.flaticon.com/authors/becris and www.freepik.com for the icons.

Figure 1.1: Illustration of Methods and Chapter Summaries

Chapter 2

Literature Review

2.1 Introduction

This research crosses a number of areas: blogs and blogging; blogging by a specific community (the Edu-blogosphere); communities of practice; document classification; dimensionality reduction; and methods appropriate for analysing the content of a large set of documents. To make this easier to navigate, it has been divided into two sections: the first covers research that relates to teachers and the ways in which they make their voices heard; the second with the technicalities of harvesting, cleaning and classifying data.

The first section includes a thorough review of the Edu-blogosphere with a focus on what the previous research reveals about the content of Edu-professionals' blogs, and the methods used to extract the information. This is important to the Methodology chapter, as establishing what is already known provides part of the framework necessary to answer Research Questions 2 and 3 (change of content over time and correlation with wider events in Education). The methods used to discover the content are important because this research is answering a series of linked Research Questions at scale; manual methods from the Social Sciences are infeasible.

The second section is not intended to be an exhaustive review of all research into the reliability and application of a range of classification and/or clustering algorithms from Computer Science. Rather, it is intended to provide an overview of the way the appropriate algorithms work, justifications for the choices that will be made (with more detail in the methodology) and will discuss some of the key research papers relating to document clustering and classification.

2.2 Part 1: Teacher Voice in Education Policy, Teaching Communities and the Edu-Blogsphere

2.2.1 The Edu-Blogosphere

It is easy to forget that the ability to set up a blog and publish - as well as linking to other bloggers as a way of building a community *and* promoting your own blog - has only been readily and easily available over the last ten years or so. Indeed, the first set of data collected for this research dates from 2004, a year after Wordpress was launched. No doubt there are older blogs somewhere on the web, and of course many teachers have always kept diaries and other written accounts of the things that interest or frustrate them, or pique their professional curiosity. There will also be many blogs contained within 'blog service' sites as defined by Schmidt (2007), but this research has focused on 'blog script' (ibid) packages which, while not always 'owned' by the blogger, offer a greater degree of freedom in terms of appearance and construction.

The reason for focusing on these 'independent' platforms is precisely *because* they are independent. They are defined by Lantz-Andersson et al. (2018) as 'informally developed' online teacher communities that form 'from the bottom up'. A 'blog service' site is self-selecting in so far as members have chosen to set up an account, and are arguably constrained by the socio-technical 'rules'. They are generally formed 'from the top down' (Lantz-Andersson et al., 2018). Edu-bloggers who set up their own blogs may write in order to find and engage with an audience, or not. Compared with formally-organised sites, the limited interactivity offered by blogging is seen as an advantage by some (Schiano et al., 2004; Loving et al., 2007). Furthermore, independent or blog-service sites are all in the public domain, which makes it possible to harvest the necessary data and analyse it at scale.

For those researchers looking to provide some insight into the *content* of blogs, there are different approaches that can be taken, and different theoretical frameworks that can be applied. Some researchers have considered blogs as ways for the author to construct a professional identity (Luehmann, 2008a; Kirkup, 2010); a variety of other approaches are neatly summarised in two recent reviews of the literature space by Macia and Garcia (2016, p.296) and Lantz-Andersson et al. (2018, p.307). They include 'Communities of Practice' (discussed below), and 'Sociocultural theory'. Both of these papers review 'informal' and 'formally-organised and informally-developed' online communities, which include a wide range of online spaces such as discussion forums, wikis, email groups and the generic sites such as Twitter and Facebook. The papers reviewed by Macia and Garcia, which selected papers published after 2009, barely refer to blogging except as one of a range of tools available to the online community. The slightly later review by Lantz-Andersson et al. discusses blogging in a little more detail, but only one of the selected studies has the word 'blog' in the title. Of the 16 papers collected for the purposes of this study, 10 use a form of the word 'blog'. More recently, the focus of research for COPs have focused on sites like Twitter.

The methods used to research online communities are reported by both Macia and Garcia Macia and Garcia (2016) and Lantz-Andersson et al. Lantz-Andersson et al. (2018) as mainly qualitative: 11 out of 23 for Macia and Garcia; 13 out of 28 (for informally developed online communities) by Lantz-Andersson et al.. The remainder were of mixed method, although Lantz-Andersson et al. identify 8 using quantitative evaluations which include "e.g. social network analysis techniques". None of the studies using this method referred to blogs; many of them were focused on Twitter. Given that, in general, the sample sizes were quite small, it's not surprising that the qualitative method was used so often. Case studies, surveys and observations were the the most reported by Lantz-Andersson et al. (20 out of a total of a total of 27 studies into informally developed online communities). Of the 24 studies selected by Macia and Garcia, 10 were centred on the United States.

To summarise the two review papers by Macia and Garcia and Lantz-Andersson et al. in terms of the parameters of this study, there is no research that uses a quantitative methodology drawn from machine learning to survey and explore the content of many thousands of Edu-blog posts contained in 'informally developed' or 'blog script' sites, nor any research that aims to discover any themes in the content that may appear (and recede) over time. Both papers do, though, record summaries of topics of discussion contained in the studies they selected, which will be explored in more detail after reviewing some of the papers not covered by the authors.

One of the first pieces of research carried out using qualitative research methods was (Ray and Hocutt, 2006). Based on a small sample of 16 teachers, the authors nevertheless identified many of the reasons teachers blog which have been referred to again by subsequent studies. They placed their findings under two headings: 'reflective practice' and 'collaboration and social interaction', which included other reasons such as sharing ideas, concerns and 'venting of frustrations' as well as overcoming feelings of isolation. 'Reflective practice' also included an example of concern for changes in the curriculum. The authors also gave a further example of professional concern in the 'collaboration and social interaction' section where they highlighted a bloggers unease with expressing their feelings about teaching in a public forum.

Interestingly, 15 of the 16 bloggers were already blogging anonymously. In an article by Kirby and Cameron (2008), which focuses on blogs written by University academics, the authors identify similar reasons: reflecting on the practices of academia, and using blogs as 'virtual soapboxes'. Again, some bloggers preferred to remain anonymous.

A paper by Luehmann (2008a) studies just one Edu-blogger, framed as 'professional identity development', which is arguably at the root of every teachers' blog. While this research is concerned to categorise blog *content* and link it with events external to schools themselves, it would be reasonable to say that teachers do not blog publicly in response to external stimuli without considering the construction of their professional identity, especially when they have blogged over the long term and are aware that they are part of a community.

The table produced by Luehmann (2008b) on page 297 (the same paper as Luehmann (2008a), re-worked for a different journal), based on research into teacher training, includes many of the topics expressed in blogs written by educators, such as reflective practice; positioning (framed as "P5: Consideration and integration of an *expert voice*"); and professional development ("P4: *Studying practice* in a way that is connected to, yet removed from, content-specific daily practice"). The subject of Luehmann's research is a 'middle school science teacher' from the United States. A later paper (Kirkup, 2010) focuses on blogging in academia, and argues that the blogs studied were not "...performances of a teaching identity" (Kirkup, 2010, p.81), but refers to them as building 'intellectual identity', perhaps reflecting the different positions of the bloggers who are both teachers *and* researchers.

In 2009, the one example where quantitative approaches have been used was published. Galyardt et.al (2009) used LDA[1] to analyse the substance of conversation in Classroom 2.0[2]. LDA is is generative[3] statistical model from natural language processing that is used to discover the topics contained in a document (and is discussed in more detail in section 2.5.3). Galyardt et al. were researching the discussion forums in Classroom 2.0, where conversations would be expected to range across different subjects. However, only the forum and wall posts were subject the topic modelling; the blog section of the site was ignored. The topics identified were site-specific - the site is largely concerned with providing IT-focused support to teachers - so were of limited interest to the context of this research. However, the method is relevant, and this paper is mentioned again in the section of this chapter that reviews research on topic modelling. Classroom 2.0 is a blog service site. Other research papers that use similar sites include Hou et al. (2010); Duncan-Howell (2010); Booth and Kellogg (2014); Robson (2016) and Kelly and Antonio (2016). Only Tour (2017) does not use a specific blogging site as a source of data, although this is because she is researching the personal learning networks (PLNs) of three participants. Kiersten Greene (2013a; 2016) is another researcher that does not draw her data from a constructed site, but her research focus is rather different in that she is concerned with exposing the "...absence of teachers' voice in the policymaking process"(Greene, 2013b).

Hur and Brush (2009) researched why teachers participate in online, blog-service communities, although only one of the sites hosts blogs. The reasons given include 'sharing

[1] Latent Dirichlet Allocation

[2] http://www.classroom20.com/

[3] The algorithm works on the basis that there is an unseen structure among the documents, which it sets out to reveal.

emotions', 'combating isolation', 'exploring ideas' etc. Duncan-Howell (2010) studied a similarly constructed sites, and focused specifically on the use of such sites to aid professional learning and development. The authors also identified 'exploring ideas' as well as 'sharing resources'. Writing a year later Hou et al. (2010, 2011) found similar reasons for the teachers in their study to blog, although this time their subjects were based in Taiwan. When researching *why* teachers blog, there are clear similarities between the US, the UK and Taiwan.

We can sum up the reasons for blogging by educators from these and other research papers as follows:

- reflective practice (Ray and Hocutt, 2006)
- sharing resources (Galyardt et al., 2009; Hur and Brush, 2009; Duncan-Howell, 2010; Robson, 2016);
- sharing ideas and collaboration (Ray and Hocutt, 2006; Galyardt et al., 2009; Hur and Brush, 2009; Hou et al., 2010; Duncan-Howell, 2010; Booth and Kellogg, 2014; Robson, 2016);
- building a sense of connection (and tackling issues of isolation), also sharing emotions (Hur and Brush, 2009; Hou et al., 2010; Duncan-Howell, 2010; Booth and Kellogg, 2014; Robson, 2016);
- 'virtual soapboxing' (Kirby and Cameron, 2008; Robson, 2017);
- the giving and receiving of support (Luehmann, 2008a; Hou et al., 2010; Duncan-Howell, 2010; Booth and Kellogg, 2014);
- professional concern (including 'venting' of frustrations) (Ray and Hocutt, 2006; Hou et al., 2010; Robson, 2016);
- professional development (Hou et al., 2010; Luehmann, 2008a)
- 'positioning' (the teacher positions themselves in the community i.e. as an expert practitioner or possessor of extensive subject knowledge) (Luehmann, 2008a; Duncan-Howell, 2010; Robson, 2016).

From here, it is now possible to draw from the literature some clear categories of Edublog post i.e.

1. Continuous Professional Development (CPD) / Training / Advice (including the giving and receiving of support in the form of advice);
2. Positioning;
3. Professional concern (venting of frustrations is included in 'soapboxing');
4. Reflective practice (not as part of CPD);

5. (sharing) Resources (and ideas);
6. Soapboxing (sometimes sharing emotions);
7. Other.

CPD is a formal requirement for all teachers in England, although only schools still under the control of the local Education authority are required to make five days available for this *in addition to* the the 190 days teachers are expected to be in front of students (sometimes also known as 'contact' days). This may include an element of 'reflective practice', although it is more likely to focus on areas such as behaviour management, safeguarding issues etc. Reflective practice is an explicit part of teacher training, and some teachers continue this practice, communicating their experience through blog posts. While the less formal idea of 'the giving and receiving of support' might not seem to be part of CPD, as a teacher I would consider advice of this kind from a fellow teacher as a form of mentoring, and as such a 'soft' form of CPD.

'Soapboxing' has been separated from 'professional concern' because the tone of the posts, communicated through the use of language and syntax, is closer to the 'venting of frustrations' identified by Ray and Hocutt (2006). Examples clarifying both of these are presented in Chapter 3, section 3.5.2.

The only category from the existing research *not* carried forward to become one of the categories that will be used to analyse the content of the blog posts gathered for this research is 'building a sense of connection'. In a sense, every communication humans have between themselves, whether online via blogs or microblogs, or face-to-face contact, is a way of building connections and combating isolation. Where the availability of social media has become invaluable to some teachers is that it affords an opportunity to make connections *at any time*. It is not unusual for teachers to rarely have time to visit the staffroom at break of lunch time: being able to access social media at home can often be the only time a teacher might connect with other professionals.

2.2.2 Education Policy

As mentioned previously, it was as a result of Government policy that the subject of the influence and impact of Edu-blogs became apparent. The method by which government policy is formed and enacted is interesting in the light of the alleged lack of teacher input,and there is some evidence to suggest that the 'policy cycle' discussed below has changed.

Education policy is the responsibility of the Secretary of State for Education. New policies come about as the result of manifesto pledges being enacted when a party forms a government; when civil servants identify a problem that needs addressing, either because of an inherent flaw in the original legislation, or a need to update policy in the light of new regulations elsewhere; and when a new minister is appointed as Secretary of State, and sets about putting their personal stamp on the department. As well as the link between Edu-community blogging and Ofsted's about turn, this research is also concerned with comparing the topics raised and discussed by the Edu-community over time and possibly in response to key policy changes in Education instigated by the government. In particular, the 'difficulties' with Ofsted referred to at the very beginning of this chapter were arguably as a direct result of a short-circuiting of the policy process by Michael Gove, although Sam Freedman, an Edu-blogger himself and at one time an advisor to Gove, is quoted by the Guardian as saying " "Policy development is much easier to do in a thinktank, [...] Parliamentarians and their staff don't have much capacity. There is too much day-to-day nonsense going on." "[4].

Thinktanks are organisations that research particular issues, and advocate policy change in the light of the research. Generally, they are non-profit organisations, although questions have been raised as to their sources of funding, which do not need to be disclosed and limit transparency. Some, like the Policy Exchange, are registered charities. The Policy Exchange[5] was set up in 2002 by a group which included Michael Gove[6], so it seems unsurprising that much of Gove's thinking was done before he took office in 2010, and with the assistance of the Policy Exchange. This was confirmed by Freeman in an article that appeared in *The Guardian* newspaper in 2013[7], which sought to shine some light on to the process of consultation under the headline "Who is really behind Michael Gove's big education ideas?". Other influential 'think tanks' include *Civitas*[8], the Education Policy Institute[9] and *Reform*[10]. Education think tanks, and some idea of their political standing, are listed on the 'Teacher Toolkit' web site[11].

The 'policy development' referred to by Freedman is known in the civil service as the 'policy cycle'. Chapter 3 of 'The Green Book' (Scholar, 2018) presents the elements of the policy cycle, although 'Policy Making In The Real World' (Hallsworth et al., 2011) makes it clear that this isn't always followed. The ideal policy cycle is shown in figure 2.1, but there is often a gap between theory and practice, sometimes for legitimate reason. Hallsworth et al. point out in the executive summary, this may be as a result of "unrealistic models" or lack of support to implement the theory. Certainly, a minister can rush through a Bill, which later becomes a policy when the Bill receives Royal Assent in parliament and becomes an Act. When Michael Gove was appointed Secretary of State

[4] https://www.theguardian.com/education/2013/dec/03/michael-gove-education-dominic-cummings-policies-oxbridge

[5] https://policyexchange.org.uk/

[6] https://en.wikipedia.org/wiki/Policy_Exchange

[7] https://www.theguardian.com/education/2013/dec/03/michael-gove-education-dominic-cummings-policies-oxbridge

[8] http://www.civitas.org.uk/

[9] https://epi.org.uk/

[10] http://www.reform.uk/

[11] https://www.teachertoolkit.co.uk/2016/11/13/think-tank-list/

for Education in May 2010, his Academies Act received Royal Assent on 27th July that same year (Gillard, 2015).

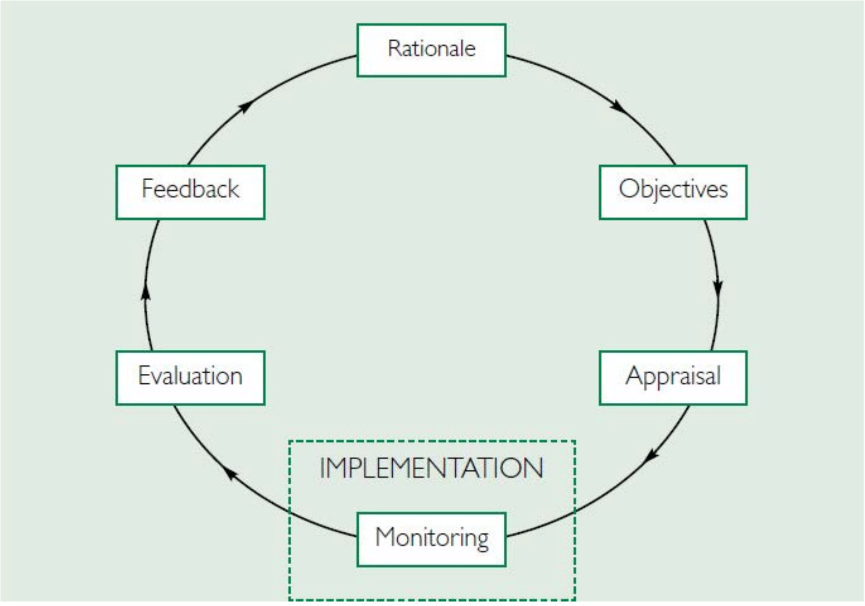


Figure 2.1: The ROAMEF policy cycle.

Martin Stanley, a former senior civil servant, set up a Web site in 2000[12] to provide advice and useful information to new civil servants. In it, he suggests a number of "discernible key stages" as a new policy makes its way from inception to enactment. There are:

- identifying a problem, and then researching it;
- consultation;
- analysis;
- presentation;
- navigating Whitehall to obtain agreement.

Of course, this assumes that the motivation from the policy arises *from* the civil service, which isn't always the case, another reason why Hallsworth et al. (2011) suggest the ROAMEF cycle (see Figure 2.1) fails. Certainly, Gove had little time to arrange consultations *before* presenting the Academies Bill, which may have been his intention. Sam Freedman, formerly an advisor to Gove and now chief executive of Ark[13] schools' Education Partnerships Group[14] stated in an article published in the Guardian[15] that Gove *had* done all his thinking and preparation in the Policy Exchange thinktank before

[12] https://www.civilservant.org.uk/policy_making-homepage.html
[13] http://arkonline.org/
[14] https://epg.org.uk/
[15] https://www.theguardian.com/education/2013/dec/03/michael-gove-education-dominic-cummings-policies-oxbridge

his appointment. In any event, the impact of this and subsequent changes to Education impacted the blogosphere, as we shall see.

Bloggers do not blog in isolation, and although this research has not collected comments left beneath blog posts, or linked bloggers with Twitter (many are active members of the Edu-community on Twitter) there can be no doubt that underlying all of these blogs is a community or communities, which may be *informal* or more formally defined as 'communities of practice' and/or 'personal learning networks'. There has been much research carried out looking at online teacher communities, and therefore a review of the literature in this area will be included.

2.2.3 Communities of Practice and Personal Learning Networks

By the late 1990s, improved access to the world wide web (WWW) led to scholars writing about *virtual communities* that had begun to appear online. The concept of the virtual community was first introduced by Rheingold (1993). He saw no difference between 'real' and 'virtual' communities. The www (or, as most people refer to it, 'the internet') was just another way in which people could connect with each other, and as a method it was absorbed and exploited along with other technology such as the telephone. Rheingold was one of the earliest users of the *The WELL*, one of the oldest communities[16], formed in the 1980s. Communities like 'The Well' would best be described as 'bulletin boards' where people could log in and join in a conversation in the appropriate 'room'. Rheingold and others (Wellman and Gulia, 1999; Wellman, 2001) focused on the relationships and social interaction facilitated by the WWW; later scholars have focused on the *content* of the conversations and posts.

The idea of communities as formal 'communities of practice' were first proposed by Lave and Wenger in 1991 (just before Rheingold started writing) as a theory of learning, in which they argued that learning happens wherever groups of people gather to exchange information. One important factor of communities of practice is that they must have a *domain* in which to work; a domain in which a group *identity* can be constructed and maintained. While some teachers and other Edu-professionals may blog with absolutely no intention of finding an audience, or becoming part of a community, most Communities of practice (CoPs) evolve naturally, although sometimes they are formed with a deliberate intent to share knowledge and enable personal and professional development. The authors' focus was on knowledge acquisition by 'apprentices' (or newcomers) learning *in situ* in a constructivist manner i.e. learning by participating in activities rather than being taught by a teacher. Learning is as much about understanding how to *behave* within the community as it is about acquiring knowledge.

[16] https://en.wikipedia.org/wiki/The_WELL

Johnson (Johnson, 2001) defined virtual communities as "...groups that use networked technologies to communicate and collaborate." (ibid, p.56). He asserts that COP have the following components that distinguish them from from 'traditional' organisations:

- different levels of expertise which are present simultaneously;
- the ability for a member to move from novice to expert within the community with complete fluidity;
- real-world tasks and communication.

Within a COP, community knowledge is greater than individual knowledge, and there is an environment of trust and safety. Johnson does not give a concrete example of an online community, although he does mention that the virtual spaces are 'mainly text based'. Given that Wordpress wasn't launched until 2003, and the advances in technology which now make it much simpler and quicker to upload a variety of content including images and video, it is fair to say that blogs often contain a great deal more than text. This is true of the current Edu-community, although the volume of text still forms a high proportion of the posts. One positive aspect of online space mentioned by Johnson (and discussed in the literature reviewed by him) is the lack of 'group norms' governing behaviour and communication. Introverts are not disadvantaged, and there is no 'voice of authority' (literally or metaphorically) to impose 'rules' and risk stifling debate.

Personal learning networks (Trust et al., 2016; Tour, 2017) differ from CoPs in that, to borrow a term from media studies, it can be regarded as a 'pull' medium. Teachers and other Edu-professionals draw on resources as and when needed, and actively increase their networks by interacting with others. Although this study does not focus on Twitter, it is worth mentioning that it is highly likely that Twitter has been instrumental in bringing blogging communities together and helping to create PLNs and CoPs. In 2011, Education Week mentioned Twitter as an example of a personal learning network (PLN)[17] that was helping teachers find ideas, resources, and support. A couple of years later, a paper (Cho et al., 2013) focusing exclusively on Twitter (and the inspiration for an article by the Times Educational supplement (TES)[18]) examined the "...service's unique limitations [that] could also be a source of strength and innovation among teacher communities online..." (ibid, p.47).

Two recent papers confirm that this is indeed the case (Carpenter and Krutka, 2017; Visser et al., 2014). Visser et al.(2014) confirm that 'blogging matters' are important for professional development. "The "blogging matters" subtheme highlights one of the most common ways professional development is created and shared via Twitter: education blogs" (ibid. p.404-405). In an earlier paper, Carpenter and Krutka (2014) also noted

[17] http://www.edweek.org/ew/articles/2011/10/26/09edtech-network.h31.html

[18] https://www.tes.com/news/school-news/breaking-news/why-twitter-could-hold-secret-better-cpd-0

that "through synchronous chats or asynchronous tweeting, educators contribute and discuss ideas, as well as sharing and acquiring resources by tweeting links to education-related articles, blogs..." (Carpenter and Krutka, 2014, p.419) further confirming the role Twitter plays in the establishment of the Edu-blogging community. Although the participants in both studies were based in the US, there's no reason to suppose that the picture is any different for social media use in the UK, as a recent article in the Guardian suggests[19].

Professional learning networks (PLNs, sometimes referred to as *personal* LNs when discussing non-professional communities and groups i.e. students) is a similar concept to COPs, although PLNs frequently lack the requirements of a specific shared identity through the use of a domain, or a sense of community (Wesely, 2013, p.307). Twitter has been a useful conduit for teachers seeking to develop their own PLNs, as evidenced by Davis (2015); Trust et al. (2016); Krutka et al. (2017) and Visser et al. (2014). PLNs allow teachers collaborate; share resources; and find links to articles, research papers and blogs that will assist them in their own CPD. All the other 'soft' benefits remain e.g. combating isolation, building a sense of community, emotional and professional support etc. but an individual can build a PLN without any of this. PLNs do not have to be part of a network where conversation takes place, although in practice this is usually the case.

Communities of Practice and Personal Learning Networks are included here because we can see blogs written by Edu-professionals as evidence of these communities, formal or informal. Both are also evidence of 'teacher voice', regardless of whether the voices are internally facing i.e. they are written primarily as a way of speaking to other teachers, or meant to be read as both internally and externally facing i.e. to reach those people who are regarded as being 'outside' the community such as government ministers. Having looked at the development of Education policy - a place where teacher voice appears to have been excluded - to communities where teacher voice is expressed, the next section examines the locus of the voice: blogs.

2.3 Part 2: Content Analysis and Machine Learning

The Web as a technology has facilitated the creation of a huge corpus of online text such as news; encyclopaedias such as Wikipedia; comments and reviews on retail sites; and posts on social media sites including blogs. These can run into tens and often hundreds of thousands in number. It is now impracticable to gather and research this data *at scale* without tools from computer science such as *document clustering* (Jain, 2010). All of these sources (news, blogs, tweets from Twitter etc.) are commonly referred to as 'documents'.

[19] https://www.theguardian.com/teacher-network/2017/apr/20/teachers-on-twitter-why-join-get-started-social-media

Document classification is a broad area of research that necessarily includes pre-classification steps such as data cleaning, dimensionality reduction, and ways of transforming the lexicon *before* any form of classification using an algorithm can begin. Furthermore, research papers and journal articles from computer science tend to focus on the performance and evaluation of one of more algorithms, whereas papers and articles researching the *content* of documents using machine learning tend to spend less time discussing the choice of algorithm and its parameters, and more on the results (Agrawal et al., 2018). This is of course understandable when one takes into account the disciplines from which the research arises. This is a classic interdisciplinary problem which arises from the choice of methodology or methodologies used, and the main discipline from which the research originates. Therefore, papers and articles that refer to *applying* machine learning to document classification and/or clustering (the difference is explained below) will be included; technical information regarding the way the algorithms address the data will be included for information.

Once a corpus is ready for classification, a number of algorithmic choices confront the researcher, one or two of which may be clear contenders or more usually a range may need to be tested before a suitable model is found. As mentioned previously, once a collection of documents becomes large, discovering the *content* of the collection becomes difficult (Hopkins and King, 2010). Much potentially interesting and useful information is obscured. Various methods have been developed in order to discover the content, all of which rely on the transformation of the text to numerical data, and then applying various measurements to the data. In order to reduce the computational overhead, one of the first steps is to try and reduce the total number of discrete words within the documents. One of the first steps is to consider removing words that add no value to a text, for example words such as ‘and’ or ‘but’; these are commonly referred to as *stopwords.*

2.3.1 Stopword Removal

High dimensionality is a recurring problem with data. Each data point becomes a coordinate in space, or a node in a network graph; a word in a document will become a data point; each document may contain many hundreds or thousands of words; a corpus may contain many hundreds or thousands of documents. While there are algorithms that can mathematically reduce the dimensionality of a set of data, the first step is usually to find appropriate ways of reducing the volume of data *before* processing.

Text-based data can be reduced by removing certain words. There are a few words which occur with high frequency, but carry little value - words like “I”, “a” and “the”. In 1966, G.K.Zipf 1966 proposed that the most frequently used word in the English language is used twice as often as the second most frequent word, three times as often as the third ranked word and so on, a power law that has become known as ‘Zipf’s Law’. It has been shown to be generally true of other languages as well, and in a variety of corpora, although it might be more accurate to refer to a 'near-Zipfian' distribution (Piantadosi, 2014; Moreno-Sánchez et al., 2016).

From the point of view of information retrieval and language analysis, many of these words are regarded as 'noise'. While there is no official standard list, language processing algorithms generally rely on a list provided by the NLTK, a 'natural language tool kit' for Python, which was initially based on early work by Van Rijsbergen (Van Rijsbergen, 1979), and builds on Zipf's Law, and has subsequently been expanded. The stopword list used in this paper is provided by scikit-Learn which uses a list developed by the University of Glasgow[20] (Lo et al., 2005). This list, and others, were researched by Nothman et al. (2018), who concluded that stopword lists should be adapted during processing. It is entirely possible that additional words could be removed if they appear to add little value to the overall understanding of the data.

The removal of frequently used words is an important first step in the categorisation and/or classification of documents (Silva and Ribeiro, 2003; Lo et al., 2005). As well as 'noise' words, Luhn (1960) proposed that the most frequently used words and the *least* frequently used words could be removed. Yang and Pedersen (1997) demonstrated that rare terms can also be 'noise', and aggressive removal had little effect on the accuracy of classification (using a 'nearest neighbour' classifier), although results varied depending on the algorithm used for feature construction. While some consensus has arisen among researchers studying document classification that the stop word list provided by NLTK is one necessary step to improve the validity of the results from any subsequent analysis, removing *additional* words is very much dependent on the researcher and the aims of the research being carried out (Blanchard, 2007, Section 3.4.4).

2.3.1.1 Stemming or Lemmatizing

The next step to consider is whether to stem or lemmatize the remaining lexicon. The most popular method is 'stemming', where a word is reduced to its stem or root form. Thus, 'teaching', 'taught' or 'teacher' would be 'teach'. There are several different algorithms available to stem words, but the most popular is the 'snowball stemmer' created by Porter (2001), building on his original stemmer (Porter, 1980). Once a stemmer has been applied, the contextual meaning of the word is no longer important - in effect, a document or corpus will be reduced to a 'bag of words'.

In contrast, lemmatizing removes inflectional word endings, but returns the base or dictionary form of the word, or *lemma*. Therefore, the algorithm must be able to identify the intended part of speech and meaning of a word in a sentence. This can be challenging, not least because any language in day-to-day use expands lexically as well

[20] http://ir.dcs.gla.ac.uk/resource/linguistic_utils/

grammatically. WordNet, developed by Princeton University[21] is probably the most popular lemmatizer, although it has not been updated for some time. Some attempt has been made to compare the efficiency of each method for the purposes of testing queries executed on a database (Balakrishnan and Ethel, 2014); however as they note even though lemmatizing the corpus used by them produced slightly better results, a different corpus might produce markedly different results (and employed the LemmaGen lemmatizer[22]).

Most research into document classification uses a stemmer as part of the methodology, or develops a bespoke lemmatizer where, for example, the language is specialised, such as biomedical text (Liu et al., 2012). The other advantage of stemming a corpus is that the total number of unique terms is reduced, therefore reducing the size of the document space.

2.3.2 Algorithms for Exploration or Confirmation

When attempting to classify documents, we can think of ‘big data’ as a collection of documents that is too large to label by hand. However, it is possible to search the web for a selection of documents that match certain key words or phrases, plus a further list of documents that probably match the search terms. This is simple example in use every day when we use an online search engine such as Google, of how machine learning can cluster documents by similarity, and we can apply similar techniques to any large corpus.

Jain (2010) introduces the approaches to finding patterns in documents as ‘*exploratory* or descriptive’ in that “...the investigator does not have pre-specified models or hypotheses but wants to understand the general characteristics or structure of the high-dimensional data...” (Jain, 2010, p.651). Text *clustering* algorithms (Aggarwal and Zhai, 2012, Ch.4) find groups of similar documents in the corpus. They are ‘exploratory’ because they use ‘unsupervised’ methods, also described as ‘non-deterministic’ because although the algorithm will iterate through the data until a result is reached, the result may vary on subsequent runs.

An algorithm for *confirmation*, by contrast, is one that deploys a statistical approach and reports a result with a replicable degree of accuracy. Text *classification* algorithms (Aggarwal and Zhai, 2012, Ch.6) are ‘confirmatory’ or ‘probalistic’ and generally ‘supervised’ because they provide the researcher with the opportunity to build a model using labelled data, and test it on unlabelled data. Once built, the model will perform consistently across the unlabelled data, returning results that are accurate to a percentage indicated by the model at the training and testing stage.

[21] https://wordnet.princeton.edu/

[22] http://lemmatise.ijs.si/

There are a range of algorithms of each type. Before discussing them, however, the text data has to be transformed so that the words can be represented by numbers. The most common methods are either *term frequency* (TF - a count of all the words used in the corpus, followed by a count of the number of times each word is used in each document); or *term frequency* normalised by the *inverse document frequency* (TFIDF). This is explained in more detail below.

2.3.2.1 Unsupervised or Exploratory Clustering Algorithms

Perhaps the most fundamental challenge when exploring large sets of structured and unstructured data is to find some pattern - some way in which the data can be grouped or clustered. Of course, this assumes that there *is* a pattern to be discerned, some relationship or relationships between parts of the data. Assuming that there is, K-means can best be described as a clustering algorithm which is part of a clustering method of finding groups of similarity within a set of data (Jain, 2010, p.659). The danger here is that algorithms such as k-means can be run repeatedly, adjusting the parameters each time, until the researcher is satisfied with the result, resulting in a risk of 'data fishing'. This is specifically addressed in Chapter 5 with regard to the clustering algorithm used in this research.

Once either TF or TFIDF has been calculated for each document, the resulting list of numbers (the 'score' for each word) can be used to calculate where each word or document might be placed on a three-dimensional graph. This 'distance model' can be generated in different ways (this is discussed in more detail in Chapter 3). A clustering algorithm such as k-means can provide information about the underlying structure of the data by finding some aspect of the data that is similar.

A cluster "...can be defined as high density regions in the feature space separated by low density regions" (Jain, 2010, p.655), although the decision as to where the 'cut' is made between regions varies between clustering algorithms. The basic k-means algorithm generates a given number of randomly-placed starting points (the number of clusters as selected by the researcher) and iterates through all the data, clustering together the documents that are closest to each other. Because the starting points (referred to as 'seeds') are assigned randomly, the resulting clusters can be different each time the algorithm is applied. *How* different they are depends on various factors, starting with how well the data has been cleaned and the stopwords removed (see the following Chapter); whether the corpus has been stemmed or lemmatized; and the distribution of the data once. Some illustrations showing how the underlying data space might look are provided in Chapter 5, section 7.10.

Other approaches to the problem construct the data as a graph, and search for the most densely connected components, such as DBSCAN or Spectral Clustering; or 'hierarchical' clustering[23]. Agglomerative hierarchical clustering splits the data from the 'bottom up' where each data point is its own cluster, and is then paired with its nearest neighbour, with pairs of clusters merging until they form one large cluster. An example is given in Figure 2.2.

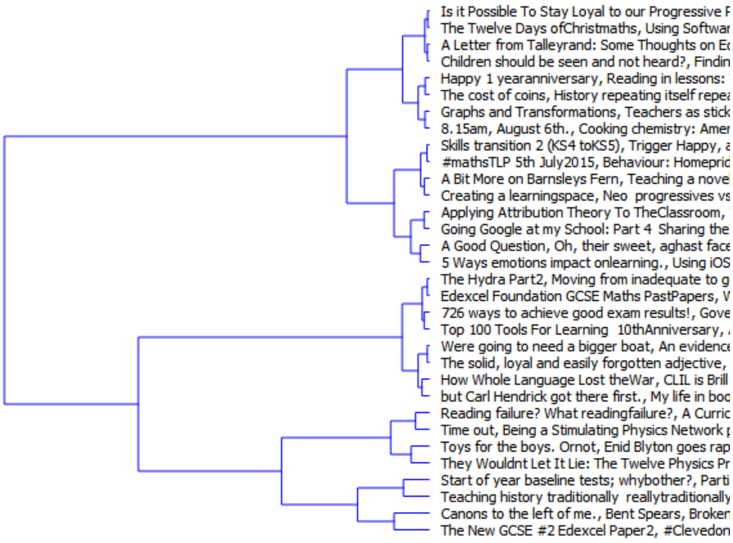


Figure 2.2: Hierarchical Agglomerative Clustering

The k-means algorithm was comprehensively reviewed by Jain in 2010, and highlights a number of drawbacks and challenges in the use of clustering alrogithms, while also acknowledging their value to data science. The main challenges are how many clusters are there in the data; which clustering method should be used; are the results valid? There is no way to accurately discover how many clusters the data contains, as a glance at the various data representations throughout the paper will make clear, as it is computationally expensive to plot a large set of data and 'have a look' prior to choosing an algorithm (and this leaves the researcher open to the ethical dangers of 'data fishing'). Indeed, as Jain (2010) says, "*...there is no best clustering algorithm*" (ibid, p.659). He notes that, when it comes to grouping data, "...the representation must go hand in hand with the end goal of the user" (ibid, p.565). Domain knowledge is essential in planning the steps that will impact the representation of the data, but this inevitably draws in some element of bias, which the researcher must constantly be aware of.

Probalistic models have also been developed, which do not assume data belongs to a particular cluster exclusively, but may be present in several. Topic modelling uses word frequency to discover the hidden semantic structures within a document. The more frequently a word is used, the more likely it is to represent a topic in the text e.g. the words 'inspection', 'visit' and 'Ofsted' appearing frequently in a document suggest at least part of the document may be discussing an Ofsted inspection. A document may

[23] http://scikit-learn.org/stable/modules/clustering.html#dbscan

contain one or more topics, with one topic being dominant over the others. An excellent illustration of how topic modelling works is shown in Figure 2.3.

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

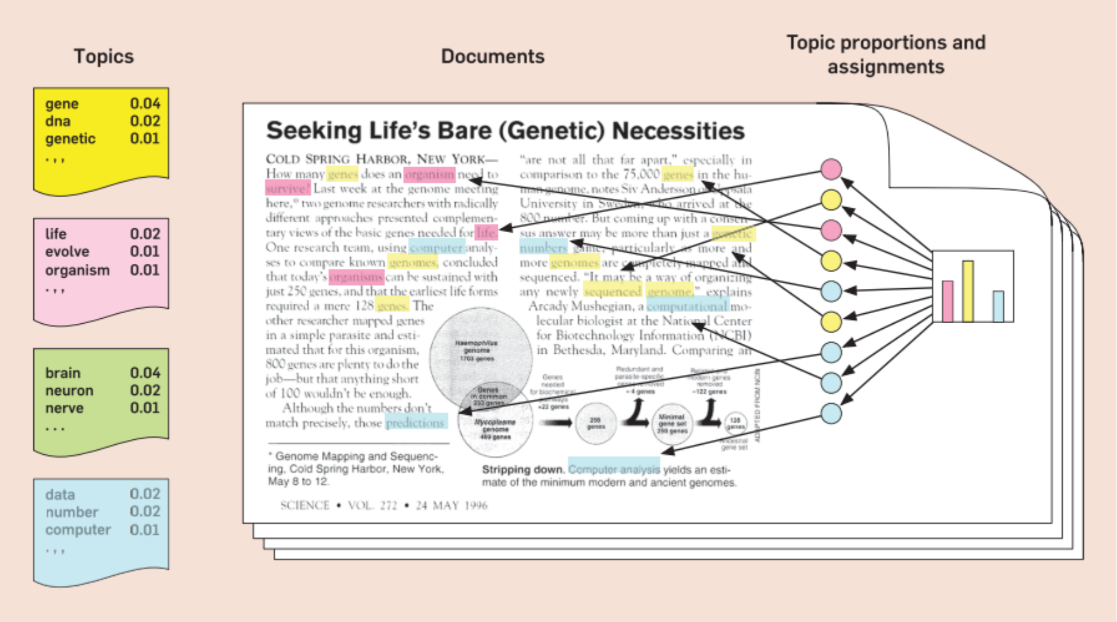


Figure 2.3: Topic Modelling (Blei et al., 2010)

Probalistic topic modelling algorithms include Latent Dirichlet Allocation (LDA), a commonly used model[24]. The assumption behind LDA is that all the documents in the corpus "...share the same set of topics, but each document exhibits those topics in different proportion" (Blei et al., 2010). However, one drawback of topic modelling (and LDA) is that, just as in k-means, the number of topics must be specified in advance. Secondly, if the data is shuffled i.e. the rows are randomised, different topics are generated (Agrawal et al., 2018).

The principal author of LDA first published a paper in 2003 (Blei et al., 2003) in which the authors described how the algorithm worked, and presented their results. Since then, it has been developed and is widely used by researchers to classify a large corpus. Four recent examples include newspaper coverage (DiMaggio et al., 2013; Jacobi et al., 2015), and academic discourse (Murakami et al., 2017; Eickhoff and Wieneke, 2018). The paper by DiMaggio et al. (2013) is particularly interesting as it addresses a similar problem when compared with this research - discovering the content of a large corpus, and linking it with external events to assess impact, although here the *tone* of the articles in the corpus was also measured to assess the extent to which support for 'the arts' was being undermined. Each paper acknowledges the limitations of topic modelling in so far as, without expensive validation procedures post-results i.e. hand coding, it is difficult to be

[24] http://scikit-learn.org/stable/modules/decomposition.html#latent-dirichlet-allocation-lda

certain as to the accuracy of the results. However, even hand coding may be infeasible, as will be discussed in Chapter 5.

The only research into the edu-blogosphere to use machine learning was a paper published in 2009 (Galyardt et al., 2009). As mentioned in Part 1, this paper did not apply topic modelling to blog posts contained within the site (Classroom 2.0) but used posts and replies from the forum, and 'wall' postings (each member has a personal comment wall, similar in structure to Twitter). The possible range of topics within the site is probably limited because the purpose of the space is to provide support for teachers using web technologies in the classroom, although the authors still managed to establish 20 topics in the forum, and 15 from the walls. A scripted site set up to specifically encourage teachers to share resources would lead to topics such as KS1, KS2, KS3, KS4, worksheets etc. Stopwords were not removed, which the authors justified on the basis that the postings were generally short and therefore removal was demonstrated as having an adverse effect on the results.

Topic modelling using LDA does have some challenges, as researched and discussed by Agrawal et al.. As well as the drawbacks mentioned above (number of topics and re-ordering the documents), using the algorithm 'out of the box' with no tuning can also give misleading results. The paper goes on to suggest a way of improving the results of LDA, but in essence the overall argument is that the results of applying an algorithm must be interpreted with care. Equal care must be taken in the preparation of data. A mixed-methods approach involving topic modelling and human coding has been developed by Eickhoff and Wieneke (2018) which uses topic modelling as the *first step* in the topic labelling process. However, the method involves a team of coders, several steps, and just over 7,000 documents. While the authors conclude that their work has the potential to produce more valid and reliable results, they also acknowledge that domain knowledge and selecting the appropriate number of topics to begin with are crucial factors.

2.3.2.2 Supervised or Confirmatory Classification Algorithms

Sometimes, it is desirable for documents to be placed in one particular class of documents exclusively, e.g. spam email. This process is also referred to as *filtering*. One of the most important decisions is how to represent the document space, and the selection of features from the data. As with text clustering, one approach is the 'bag of words' model, the other is to treat the text as a sequence of words, or a *string*. Again, stop word removal is an important consideration. Text classification algorithms are discussed in detail in Chapter 6 of *Mining Text Data* (Aggarwal and Zhai, 2012), and much of the information in this section is drawn from there.

Classifiers are concerned with prediction, and are evaluated by their classification accuracy. There are many algorithms that can be used to build a model to classify data; Scikit-Learn lists 17 different types that can be deployed using Python[25]. Choosing the most appropriate algorithm to classify text documents is therefore not straightforward, although the research literature in the area of document classification suggests that there are three main methods that are used:

- SVM (Support Vector Machine) Classifiers that partition the data by drawing non-linear boundaries between classes;
- Beysian (Generative) Classifiers which classify text according to the underlying word features in different classes e.g. the presence of the word ‘curriculum’ in a document might suggest class *a*, the probability of the document actually belonging to class *a* increasing if the document also includes the words ‘plan’ and ‘design’ when they exist in the already-labelled class;
- Nearest Neighbour Classifiers that transform the data, via a distance model, to a graph where documents are connected to one or more others with the closest documents to the labelled class being allocated to that class.

It can be seen that, with the exception of Bayesian classifiers, the unlabelled data is classified according to some calculation of distance, just like clustering algorithms and with the same challenges. A range of different algorithms used to classify the Edu-blogosphere will be discussed in Chapter 3.

2.3.3 Semi-Supervised Learning

Semi-supervised learning is a way of ‘learning’ from partially labelled data - the data that is labelled is used to label the rest of the data. At a time where the volume of data being generated is beyond the capacity of humans to evaluate (Jain, 2010), this method is invaluable. There are two approaches: *inductive* and *transductive* learning. In Chapter 2 of Zhu and B (2009), the authors describe these as “inductive ... learning is like an in-class exam, where the questions are not known in advance, and a student needs to prepare for all possible questions; in contrast, transductive learning is like a take-home exam, where the student knows the exam questions and need not prepare beyond these” (ibid, p.12). Semi-supervised learning is either concerned with *predicting* the labels of future test data, or the labels of unlabelled data in the training sample. The first is inductive, the second transductive. Given the probable size of the corpus generated by the Edu-Blogosphere, and the infeasibility of labelling each document, the method used to answer Research Question 1 will be transductive i.e. a small sub-set of data will be labelled and used to predict the labels of the rest of the training set before being tested on the remaining data.

[25] http://scikit-learn.org/stable/supervised_learning.html#supervised-learning

2.4 Research Framework

As stated previously in Chapter 1, the data gathered for this research is from the social layer of the Web. How this is done is explained in Chapter 3, and because of the large number of URLs already identified as a starting point, and the probability that the first identified blog dates from 2004, the resulting corpus will be 'big data'. In order to analyse the content of this data - to act as the lens through which the content can be evaluated - a variety of approaches from Computer Science must be considered. The framework within which the data can be situated is provided by Schmidt (2007).

Whether the preferred definition of a 'community of practice' is formal (proposed by Lave and Wenger (1991)), or less formal i.e. emerging naturally (Johnson (2001)), both can be also be seen as 'personal / professional learning networks' where knowledge is actively sought out by members of the community (Trust et al. (2016). What is important is the formation and continuation of these communities or networks through the act of blogging. As well as providing a useful definition of the two different types of blogging software (blog services or blog script packages) Schmidt (2007) proposed an analytical framework 'as a basis for comparative and longitudinal studies'. The author suggests that "...there has been no concise and systematic formulation of an analytical model of blogging practices that can integrate the varying motives, routines, and consequences of appropriation and usage of this new communicative genre" and proposes a "...heuristic framework that is grounded both in ideas from general sociological theory and in existing blog research" (ibid, p.1411).

Schmidt (2007) connects a blogging 'episode', using specific software, to the attainment of "...specific communicative goals" (p.1411) and relates this back to sociological structuration theory proposed by Giddens (1984). Blogging episodes are "...framed [...] by three structural elements: rules, relations, and code" (p.1411). This research draws on the 'rules' which we can see as the shared expectations of a community based on generalisable knowledge. These are drawn from existing research, and listed in Section 2.2. In short, this research has collected a series of blogging episodes created using blog script packages, and applies the 'adequacy rules' from the framework to establish what the Edu-Community blogs about. It then goes on to ask some additional questions to link the research with the evidence of 'teacher voice' by mapping the changes in content focus over time, and triangulating this with events from the government and its ministers on Education.

The lens with which to view the results of the 'adequacy rules' is provided by content analysis using machine learning. Two approaches are used: the first applies a label to each of the reasons arising from the application of the adequacy rules, the labels are then attached to a small training set of examples, and the remaining blogging episodes categorised using a classifier. While this cannot answer research question 1 as it does not add anything new to 'existing blog research' in terms of content, it can provide some insight for research questions 2 and 3. The second ignores the adequacy rules, and answers research question 1 with no *a priori* assumptions, and will provide evidence to answer research questions 2 and 3 as well as new insights for question 1. Together with a timeline of events from education policy, all of which is represented in Figure 2.4.

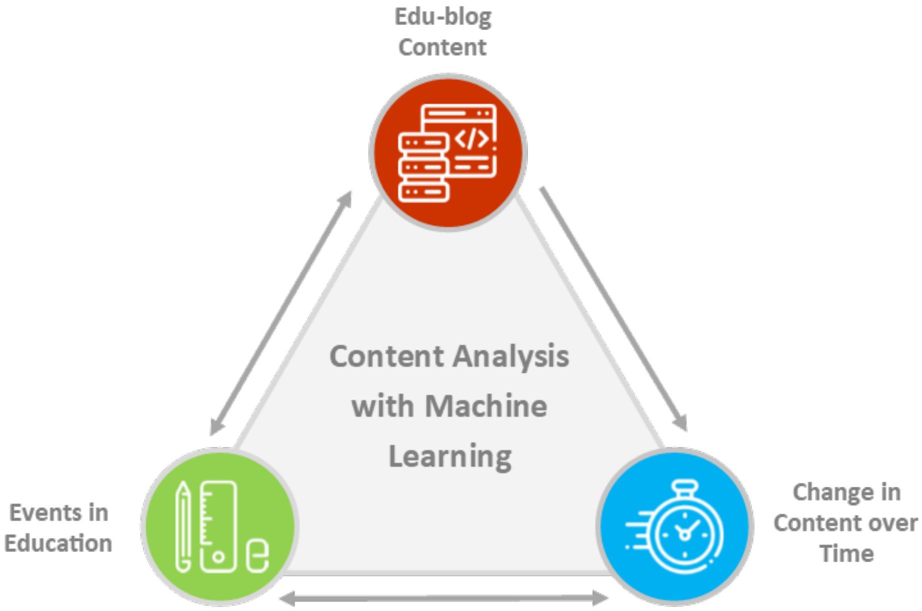

Edu-blog Content
Content Analysis with Machine Learning
Events in Education
Change in Content over Time

Figure 2.4: A Research Framework

Three data-production strands are therefore necessary to answer all three research questions in the light of the sections of the framework drawn from Schmidt (2007). While of equal importance, the first concerns the harvesting, cleaning and pre-processing of data which is presented in the following chapter, Chapter 3. The second and third are connected in that they involve a series of steps to filter out unwanted blogs before building and applying a classifier, and applying the clustering algorithm 'topic modelling'. The classifier will necessarily require an additional step - the labelling of a small training set with examples that fit the category definitions which is then used for semi-supervised learning and testing a range of different classifiers. This is discussed in Chapter 3.

Chapter 3

Methodology

3.1 Introduction

The research methodology presented here begins with the theoretical underpinnings of this type of research. Following this, the chapter is divided into three discrete sections. Part 1 deals with the way the data is gathered and dealt with *before* the content can be analysed; in short, how the data is 'transformed'. Part 2 focuses on the analysis of the blog posts themselves, and part 3 discusses how to best visualise and present the data, which must necessarily be re-formatted and summarised to fit into a two-dimensional space. Each section is represented by Figure 1.1 at the end of Chapter 1.

3.2 Positionality

Before any discussion can take place, it is important to acknowledge the role of the researcher in the process. Indeed, it could be argued that it is impossible to separate and extract the researcher from the research. The researcher brings knowledge and a framework to apply to a research question or questions, but the researcher is human and "humans actively construct their own meanings of situations..." (Cohen et al., 2007, Ch.7, p.67). Whilst the heading for this section has used the label 'positionality', (Cohen et al., 2007, Ch.14, p.310) uses the term 'reflexivity' to describe "...a self-conscious awareness of the effects that the researcher is having on the research process". The context of this is 'action research' where the researcher is also and at the same time a practitioner carrying out research in the field, but it is also true of all the research carried out here. The decisions and interpretation of results have been made in the light of the knowledge and experience of the researcher, and it is legitimate to suggest that a different researcher, even one from the same domain, might approach each stage of the research differently. The positionality - or bias - of the researcher will be acknowledged where it may have had an impact, particularly in Chapter 5.

3.3 Mixed Methods Research

Broadly speaking, research methodologies are qualitative, quantitative, or mixed methods (MM). Qualitative research has its roots in the Social Sciences - a constructivist worldview which is subjective and inductive (Creswell and Clark, 2018, Ch.2, p.38). The data is typically interviews with study participants, or similar artefacts where the focus is on people's *opinions* or *experiences*. The perspective of the researcher, who may or may not be a participant in the research, is also foregrounded in the interpretations and conclusions drawn in the study. In contrast, a quantitative approach is positivist. One single reality exists which can be observed, deduced and the result(s) stated (ibid, p.38). The researcher either rejects or fails to reject the hypothesis. Numerical and categorical data subject to statistical analysis is a common example of a quantitative methodology. A MM combines both qualitative and quantitative data in terms of collection, analysis, and interpretation(s) and/or conclusion(s). In this way, the whole of the research, from questions to design to results and conclusions, can be argued greater than the sum of its parts, and as such it has been referred to as 'the third paradigm'. This is summed up well in a recent paper (Symonds and Gorard, 2009).

The mixing of methods as the best way of addressing research questions is sometimes referred to as pragmatism i.e. selecting methods based on 'what works' (Creswell and Clark, 2018, Ch.2, p.37) and Biesta in chapter 4 of Tashakkori and Teddlie (2010). However, whatever label is applied to any aspect of a mixed methods methodology, it is not always helpful, as argued by Symonds and Gorard (2009). Examples of quantative analysis tends to focus on statistical analysis using software packages such as SPSS; or coding qualitative data thereby transforming it into quantitative using NVIVO. Cluster analysis, and graph theory to measure social networks is mentioned by Bazeley in Chapter 18 in Tashakkori and Teddlie (2010) as other ways in which data can be quantised and visualised. Small samples of data can be clustered, or networks visualised, by hand-coding. However, a social science researcher may have good reason to consider these tools to be inadequate when analysing data at scale.

The research questions articulated in this study are grounded in the social sciences. The data used to answer the first research question - what does the Edu-community blog about? - exists in the social layer of the Web (Phethean et al., 2016), and in existing research. For the first strand of data, the blog posts themselves, the answer to the *what* is arrived at predominantly through quantitative methods using algorithms from Computer Science. The second strand of data which is needed to answer the second research question - has the content of blogs by the Edu-community changed over time? - arises from a review of the existing literature which provides a series of labels for the *what*. A small training set of blog posts are then categorised with these labels, before a classifier algorithm is applied. In short, the collected data is *transformed* into quantitative data with no intervention by the researcher until the process is complete, and the results produced. At this point, the results from the chosen classifier (discussed in more detail in Chapter 4), are accepted 'as is', whereas the results from the second algorithm (which does not take into account the training set) can be subject to further parameter changes which may impact results. This second algorithm also needs the researcher to state at the outset how many topics to look for, whereas this knowledge is already available to a classifier via the results of the literature review. In short, the process of transforming and analysing the data up until the point when the results can be presented is iterative. The third research question - is there a link between any change over time, and policy-related events in education? - is answered by triangulating the data against a time line.

Using machine learning to analyse data at scale is a relatively new challenge for social scientists. The arrival of 'big data' (or at least data too big to analyse 'by hand') has shifted the focus to an almost entirely quantitative approach. However, even using machine learning is not a wholly quantitative process. Before the text data can be transformed to numerical data, decisions have to be made that are qualitative, such as the development of a bespoke stopword list that - in the case of this study - is reliant upon the domain knowledge of the researcher. Following the transformation of the data, the labelling of a training set of blog posts with category labels derived from existing research is closely related to the thematic analysis carried out by social scientists. There is no inter-code agreement step (discussed in Chapter 2), but this is the only difference. The second algorithm which ignores the training set is pre-processed in exactly the same way with regard to stopwords, but requires the researcher to consider - again based on domain knowledge - how many topics to extract from the data, and the optimum number of topics based on an iterative process of reviewing the results in the light of differing numbers of topics. Furthermore, the topics are eventually labelled with meaningful labels that may not be immediately recognisable from the list of topic words, but again the researcher has the domain knowledge to be able to make that judgement. The assumption that algorithms are quantitative, and therefore rooted in a positivist, objectivist and deductivist epistimology are not always valid. The *internal* calculations of the algorithm may be objective, but the decisions leading up the the deployment of the algorithm, the data being fed into the algorithm, the parameter adjustments made in the light of results and the interpretation of the results are based in a qualitative methodology (Agrawal et al., 2018).

The following section of this chapter discusses the decisions made regarding the data that are *not* covered in Chapter 3. The first concerns the question as to whether every blog in the original list of blogs should be included in the data set. Web Science does not demand a particular methodology or methodologies to be used - rather, it leaves the researcher free to draw upon the range available to answer the research question or questions in the way that works. Consequently, the methodology used here uses tools from Computer Science combined with a mixed methods approach that offers opportunities for qualitative analysis that are often an inherent part of the algorithmic process. What distinguishes Web Science from other disciplines is that the data is generated by a particular community exploiting the social layer of the Web, and it is the impact of this social interaction that is important to the community. Here, in this layer, is evidence of a professional community discussing any and all aspects of their professional lives, and doing so at scale. The results of running the models - a classifier and topic modelling - on the data will be presented and discussed in the following chapters.

The following section of this chapter focuses on the methods used to harvest, clean and pre-process the blog data.

3.4 Part 1: Data Transformation

In order to have confidence in the results, it is important to ensure that the data being addressed by the chosen algorithms is clean, and pre-processed so that it is in effect reduced to the most relevant aspects with regard to the research questions being asked of it. There are a number of ways in which the data can be transformed. These steps are presented in this, the first part of the methodology. The raw data for this project consists of the following:

- The date a blog entry was posted.
- The title of the blog post.
- The content of the blog post (text only).

This data will be gathered from the list of URLs obtained from Andrew Old[1] (see Chapter 1).

[1] https://teachingbattleground.wordpress.com/2015/08/12/please-help-with-the-uk-education-blogs-spreadsheet-version-12/

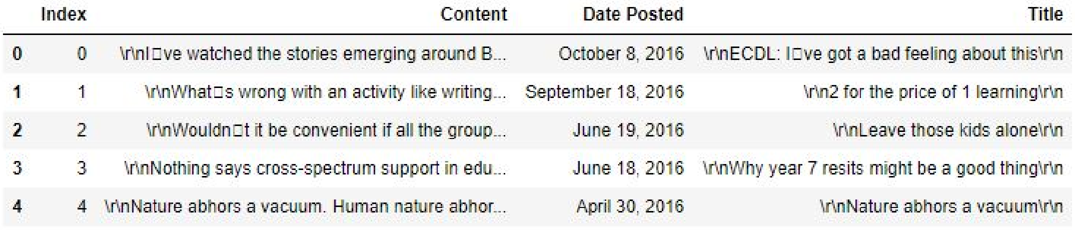
3.4.1 Data Harvesting

The first step is to gather the data. This is achieved by locating where in a web page the data is contained, and then writing code using Python to harvest the data and save it in a spreadsheet format. The module 'Beautiful Soup'[2] was used to locate the relevant data from within the html code of the web page(s), and 'Pandas'[3] to write the data to a spreadsheet. From here, the data can be accessed and Python can be used again to write the necessary code to clean and pre-process the data.

Python has an extensive library of built-in tools, as well as many thousands of packages written by third-party developers. Python can therefore be used to process numerical data, text and images, and provide data visualisation in a variety of formats. All of the analyses used here are developed using a sample set of data, and replicated on the main data.

3.4.2 Cleaning Text Data

Data rarely arrives at any spreadsheet in 'clean'. There are often characters in the text that represent spaces at the beginning of sentences, or 'carriage returns'. This is evident in Figure 3.1. Getting rid of these is quite straightforward using Python code, examples of which, written in 'Jupyter', are given below in Figure 3.2 and Figure 3.3. Jupyter is one part of a platform provided by Anaconda[4] which offers a range of different environments with which to use Python code. Jupyter has the advantage of being able to run code in batches in order to see the results of each stage, as shown in the figures referred to above.

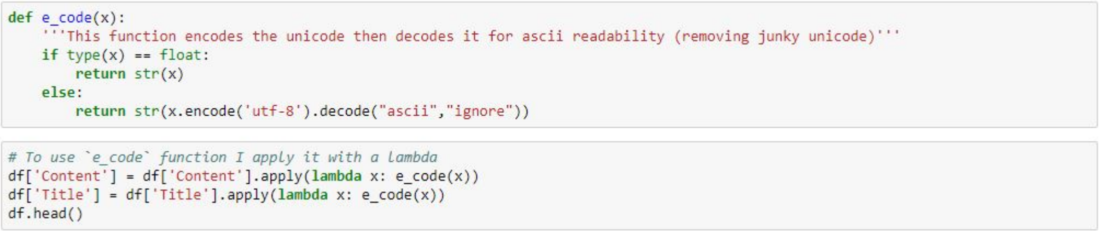


	Index	Content	Date Posted	Title
0	0	\r\nI□ve watched the stories emerging around B...	October 8, 2016	\r\nECDL: I□ve got a bad feeling about this\r\n
1	1	\r\nWhat□s wrong with an activity like writing...	September 18, 2016	\r\n2 for the price of 1 learning\r\n
2	2	\r\nWouldn□t it be convenient if all the group...	June 19, 2016	\r\nLeave those kids alone\r\n
3	3	\r\nNothing says cross-spectrum support in edu...	June 18, 2016	\r\nWhy year 7 resits might be a good thing\r\n
4	4	\r\nNature abhors a vacuum. Human nature abhor...	April 30, 2016	\r\nNature abhors a vacuum\r\n

Figure 3.1: 'Dirty' Data

However, one problem with converting the text from 'dirty' utf-8 to 'clean' ascii is that, where blog posts were written using an Apple device, the apostrophes used to denote missing letters in contractions were removed i.e. isn't became isnt. This leads to two outcomes: first of all, we have a useless word, isnt, that will clog up the data, but

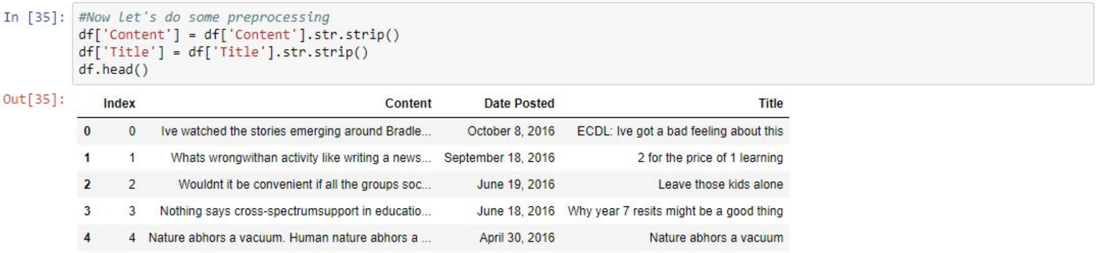
[2] https://www.crummy.com/software/BeautifulSoup/
[3] https://pandas.pydata.org/
[4] https://www.anaconda.com/



```
def e_code(x):
    '''This function encodes the unicode then decodes it for ascii readability (removing junky unicode)'''
    if type(x) == float:
        return str(x)
    else:
        return str(x.encode('utf-8').decode("ascii","ignore"))
```

```
# To use `e_code` function I apply it with a lambda
df['Content'] = df['Content'].apply(lambda x: e_code(x))
df['Title'] = df['Title'].apply(lambda x: e_code(x))
df.head()
```

Figure 3.2: Cleaning Data, Example 1



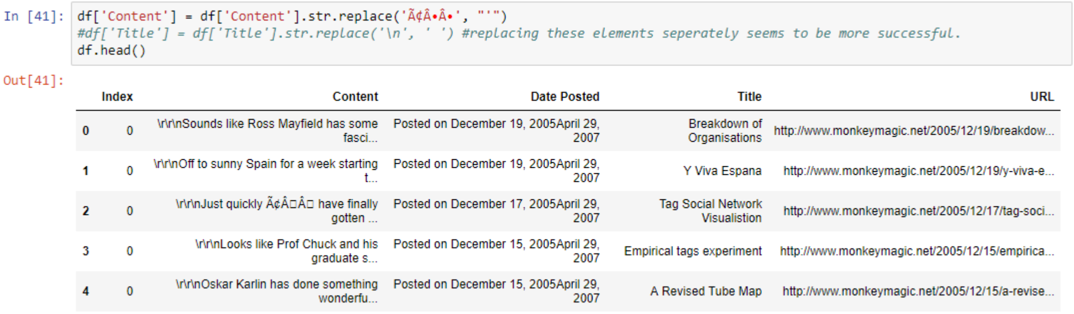
```
In [35]: #Now let's do some preprocessing
         df['Content'] = df['Content'].str.strip()
         df['Title'] = df['Title'].str.strip()
         df.head()
```

Out[35]:

	Index	Content	Date Posted	Title
0	0	Ive watched the stories emerging around Bradle...	October 8, 2016	ECDL: Ive got a bad feeling about this
1	1	Whats wrongwithan activity like writing a news...	September 18, 2016	2 for the price of 1 learning
2	2	Wouldnt it be convenient if all the groups soc...	June 19, 2016	Leave those kids alone
3	3	Nothing says cross-spectrumsupport in educatio...	June 18, 2016	Why year 7 resits might be a good thing
4	4	Nature abhors a vacuum. Human nature abhors a ...	April 30, 2016	Nature abhors a vacuum

Figure 3.3: Cleaning Data, Example 2

more importantly should the data be required for a subsequent study, the removal of the apostrophe makes it virtually impossible to expand the word to its original form 'is not'. Therefore, a preliminary step was introduced, shown in Figure 3.4



```
In [41]: df['Content'] = df['Content'].str.replace('Ã¢Â•Â•', "'")
         #df['Title'] = df['Title'].str.replace('\n', ' ') #replacing these elements seperately seems to be more successful.
         df.head()
```

Out[41]:

	Index	Content	Date Posted	Title	URL
0	0	\r\r\nSounds like Ross Mayfield has some fasci...	Posted on December 19, 2005April 29, 2007	Breakdown of Organisations	http://www.monkeymagic.net/2005/12/19/breakdow...
1	0	\r\r\nOff to sunny Spain for a week starting t...	Posted on December 19, 2005April 29, 2007	Y Viva Espana	http://www.monkeymagic.net/2005/12/19/y-viva-e...
2	0	\r\r\nJust quickly Ã¢Â□Â□ have finally gotten ...	Posted on December 17, 2005April 29, 2007	Tag Social Network Visualistion	http://www.monkeymagic.net/2005/12/17/tag-soci...
3	0	\r\r\nLooks like Prof Chuck and his graduate s...	Posted on December 15, 2005April 29, 2007	Empirical tags experiment	http://www.monkeymagic.net/2005/12/15/empirica...
4	0	\r\r\nOskar Karlin has done something wonderfu...	Posted on December 15, 2005April 29, 2007	A Revised Tube Map	http://www.monkeymagic.net/2005/12/15/a-revise...

Figure 3.4: Cleaning Data, Example 3

All contractions could then be expanded i.e. 'wasn't' becomes 'was not' etc.

The steps for the initial data clean-up can be listed as follows:

- replace code for an apostrophe with an apostrophe;
- remove unicode;
- remove whitespace from beginning and end of posts;
- remove characters indicating carriage returns, new lines etc.;
- remove duplicate posts;
- remove posts of 280 characters of less;
- remove other unwanted posts (see chapter 3);

- expand contractions.

Once the data has been cleaned, the next step is to prepare it for processing. Much of the following section will present the experimental pre-processing steps carried out on a sample of 7,225 (cleaned) blog posts. The results of these steps have been replicated on the final data. Here, the data is processed using 'Orange', an environment similar to Jupyter and provided by the platform Anaconda. Orange has the advantage of being a 'drag and drop' workspace where the underlying code is generated automatically.

All the code used throughout this research is available at the online repository GitHub[5].

3.4.3 Stopword Removal

One of the first steps after cleaning is to find ways of *reducing* the overall lexicon. Given that some complex algorithms are going to be deployed on the data, the smaller the data sets can be made, the more easily the data can be processed using a laptop computer.

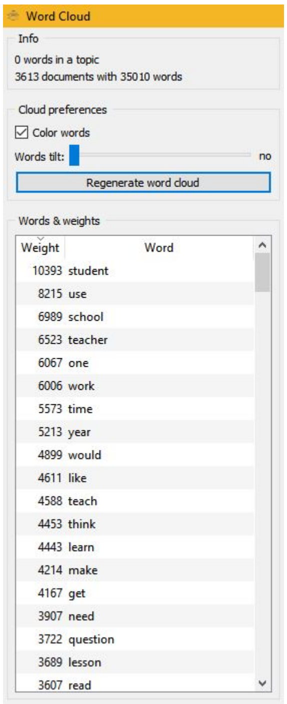
Research shows that removing some frequently-occurring words is desirable, and common practice (see 2.3.1). As well as removing a 'standard' set of frequently-occurring words, there may be others within the corpus that should be included. To discover these, the sample set was opened in Orange, shuffled, and a sample of 50% extracted. This provides a random sample which will act as a comparison with the remaining data to establish whether the results of the following analyses can be reproduced. Without removing any words *a priori*, a list of the most frequently used words was generated. The individual words in the corpus have been reduced to their word stem or root form, and thereafter referred to as 'tokens'. Figure 3.5 shows that there are a total of 35,010 unique 'tokens' (or stems) in the corpus, the most frequently used being *student, use, school, teacher, one, work, time, year, etc.*. The resulting word cloud is shown in Figure 3.6.

A wordcloud based on the remaining 50% of the sample confirms that similar words are used frequently (see Figure 3.7).

The research literature has demonstrated that removing some words is worth doing; therefore it was decided to use the list provided by Scikit-learn[6], a Python module that provides a series of machine-learning tools, some of which are used by Orange.

The following words were then added to the stop word list: *student, students, school, teach, teacher, teachers, teaching, time, year, work, use, like, make, need, think, question, lesson, lessons.* As we have seen in Chapter 2, the more frequently a word is used

[5] https://github.com/sh9g14/PhDCode
[6] http://scikit-learn.org/stable/



Info

0 words in a topic

3613 documents with 35010 words

☑ Color words

Words tilt: no

Weight	Word
10393	student
8215	use
6989	school
6523	teacher
6067	one
6006	work
5573	time
5213	year
4899	would
4611	like
4588	teach
4453	think
4443	learn
4214	make
4167	get
3907	need
3722	question
3689	lesson
3607	read

Figure 3.5: Most Frequently Used Words in Corpus Sample

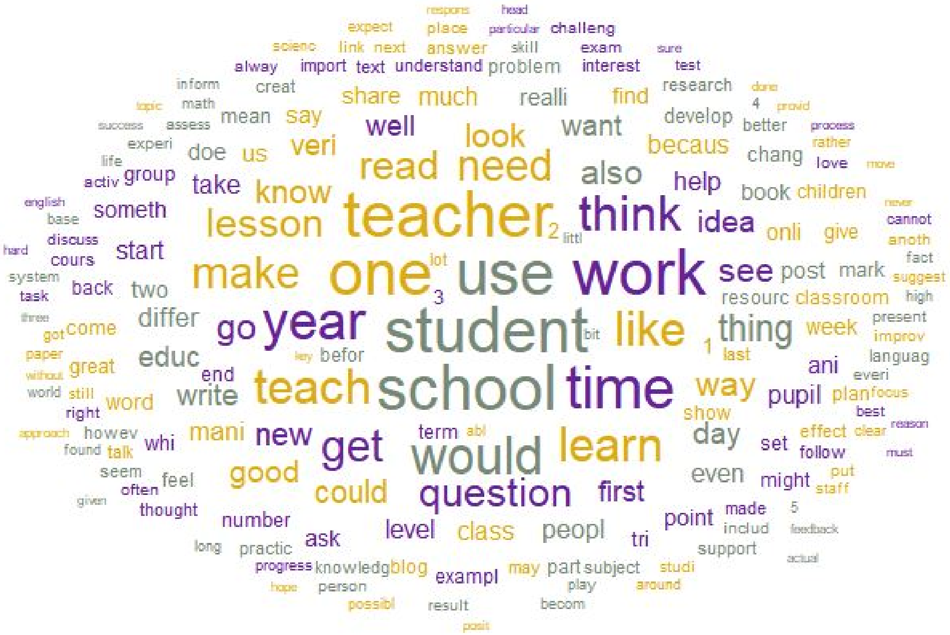


Figure 3.6: Word Cloud (50% Sample)

in a corpus, the less value it is likely to add. For example, if you were presented with a database of blogs written by teachers, and you wanted to find the blogs written about SATS, that‘s what your search term would be, possibly with some extra filtering words like ‘primary’ and ‘England’. You would know not to bother with ‘student’, ‘children’ or ‘education’ because they‘re words you‘d expect to find in pretty much everything. Those words are often referred to as ‘noise’. The models chosen to categorise the corpus

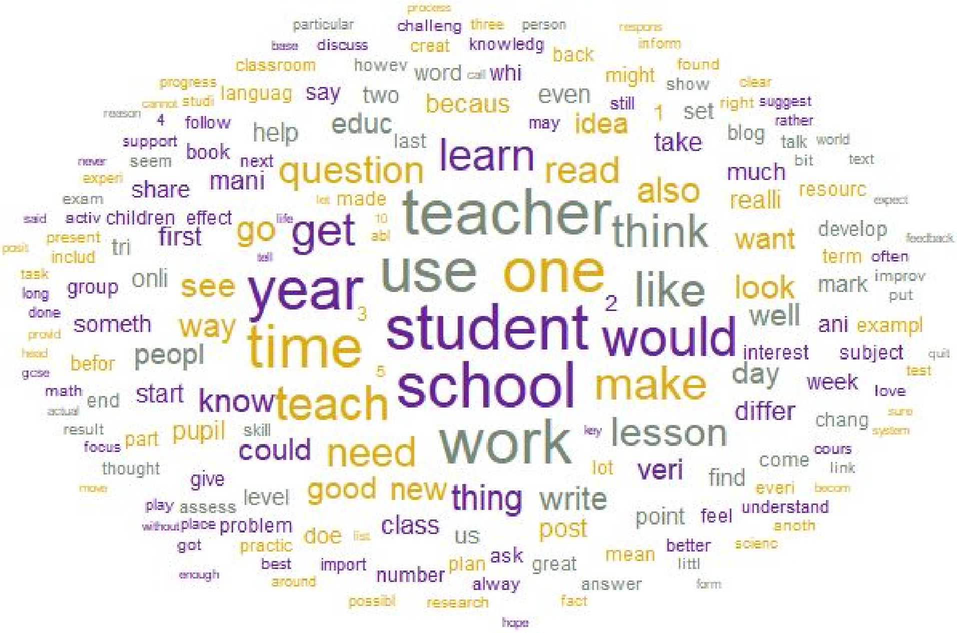


Figure 3.7: Word Cloud (Remaining 50%)

(in order to answer research question 1) were, however, tested with and without stop words to confirm there was no detrimental effect on the outcome.

In order to see the results of removing the extra stop words, another word cloud was generated (see Figure 3.8). As Figure 3.9 shows, the most frequently used words were removed to provide a clearer insight into the overall content of the corpus.

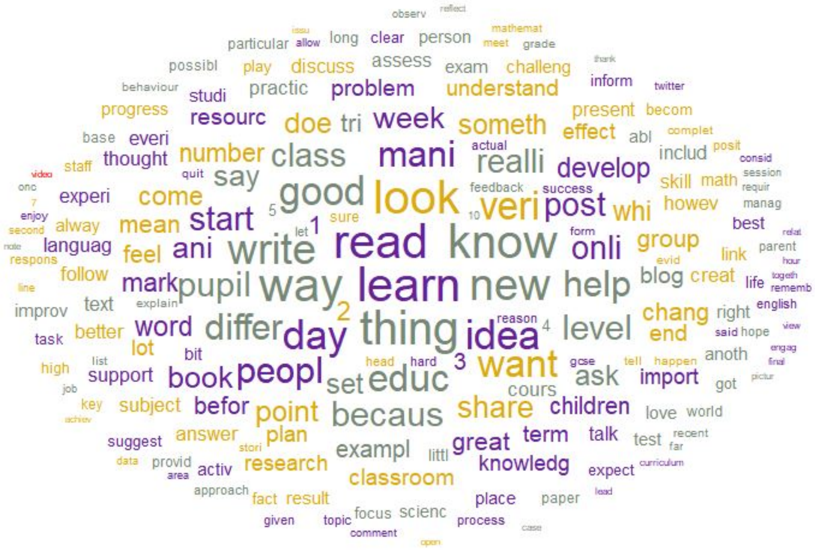
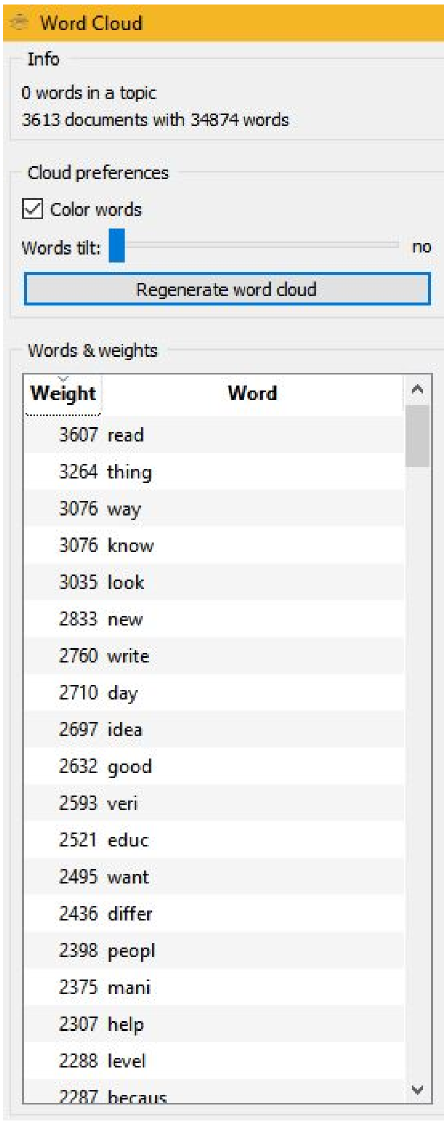


Figure 3.8: Word Cloud (Bespoke Stop Words)

The final, bespoke list of stop words is listed in Appendix A, together with the set used by NLTK and Scikit-Learn, and discussed in Chapter 2.



Info

0 words in a topic
3613 documents with 34874 words

Cloud preferences

☑ Color words

Words tilt: no

Words & weights

Weight	Word
3607	read
3264	thing
3076	way
3076	know
3035	look
2833	new
2760	write
2710	day
2697	idea
2632	good
2593	veri
2521	educ
2495	want
2436	differ
2398	peopl
2375	mani
2307	help
2288	level
2287	becaus

Figure 3.9: Most frequently used words after the removal of bespoke stop words)

3.4.4 Stem and Tokenise

Each word in the corpus is reduced to its stem form i.e. certificate/certificated/certified becomes ‘certif’; punctuation removed and each stem separated by a comma to that in effect it becomes a discrete piece of data in preparation for the analysis. This has the advantage of reducing the vocabulary of the corpus, but some nuance of meaning may be lost along with the original word.

3.4.5 Sort and Store

The entire corpus is now separated into spreadsheets by year. This is important to be able to show evidence of any change in the focus on blog posts over time, and answer Research Questions 2 and 3.

3.4.6 Summary

The blog posts have now been cleaned and pre-processed in readiness for the analysis to answer the research questions. The following section will focus on the *content analysis* of the posts using two different approaches - building a *classifier*, and building a *clustering* algorithm. The *classifier* will classify each blog post as one of the categories identified in Chapter 2; the *clustering algorithm* will ‘discover’ the topics based on the contents of the data.

Table 3.1: Research Questions

Research Question 1: What kind of subject matter, themes or issues have been discussed by the Edu-community since the arrival of the first available Edu-blog?

Research Question 2: Has the subject matter, themes or issues discussed by the Edu-Community changed since the arrival of the first available Edu-blog?

Research Question 3: Is there any link between the subjects, themes or issues discussed by the Edu-community and changes in Education policy?

3.5 Part 2: Content Analysis

The previous chapter detailed the steps taken to transform the data in preparation for content analysis. This chapter focuses on the second part of the methodology: the analysis of the blog posts themselves.

3.5.1 Defining Boundaries: Should Every Blog Be Included?

The spreadsheet curated by Andrew Old[7] is freely available for anyone to edit. Whilst this provides an excellent starting point, there have only been a couple of attempts the validate the blog URLs listed. The spreadsheet is open and freely available for anyone to edit, and it is inevitable that not all the blog URLs listed are still available, relevant or appropriate. The first task was therefore to scan through the list, open a few using a web browser, and decide on some criterion by which some URLs would be rejected (or accepted).

Whilst most of the blog URLs lead to blogs written by people involved in Education, not all of them are appropriate for this research. Some are written by bloggers from outside the UK, some of the blogs are contained within one of the teaching union web sites. Others are written by school governors, educational consultants, purveyors of technical equipment for schools, and others loosely associated with Education. In order to answer the research questions posed in Chapter 1, the blogs selected should provide a representation of the Edu-blogosphere in the UK. Whilst the list of blog URLs is 'only' around 2000 entries long, the list is too extensive for a manual check of the source of each and every one. However, during the course of writing and testing the code used to harvest the blogs it was possible to identify many blog URLs that *could* be removed. Therefore, some criteria was necessary in order to decide which blogs were relevant to this research, and would be harvested if possible.

In order for a blog to qualify for including in this research, the following conditions must apply:

[7] https://teachingbattleground.wordpress.com/2015/08/12/please-help-with-the-uk-education-blogs-spreadsheet-version-12/

1. The blog must be written in English;
2. The blog should be written by a practicing teacher (see final point). This includes lecturers in Higher Education (HE);
3. The entire blog URL, and every post contained within it, must be written by the same individual i.e. the blog URL must be 'owned' by one person who is also responsible for writing each post personally;
4. If the blog is **not** written by a practising teacher, it must be written by someone with a professional interest in Education beyond that of an Educational consultant.

The final category requires some further explanation. There are several important individuals in the Edu-community who no longer teach, perhaps one of the best known being Tom Bennett. Tom was a full-time RE teacher, but has since resigned his post in order to focus on other things, such as setting up researchED[8], and to write a report on behaviour in schools, commissioned by the government[9]. He also writes for the TES. Tom's opinions as regards matters of education, written on his own personal blog[10], I judge to be worth including, together with others in similar circumstances.

Blogs posts consisting of 280 characters or less were removed, the rationale being that Twitter now allows posts of up to 280 characters, thereby re-defining the term 'micro blog', and this also eliminates the 'hello world' and other 'this is my first blog post' posts, plus posts that simply link to something of interest with little supporting text. These posts simply aren't long enough to impart any useful information; it seemed sensible to remove them.

Finally, whilst every effort has been made to eliminate blogs that don't qualify according to my criteria, it is entirely possible that some remain.

3.5.2 Defining Categories Using Existing Research

It has already been established that using existing research to develop categories and use them to label the data i.e. using a 'classifier', is best applied to answering research questions 2 and 3. The first step is to use the categories arising from the existing research to label a small set of training data.

The categories that emerged from the literature can be characterised as follows:

1. CPD / Training / Advice;

[8] https://researched.org.uk/about/our-mission/

[9] https://www.gov.uk/government/publications/behaviour-in-schools

[10] http://behaviourguru.blogspot.com/

2. Positioning;
3. Professional concern;
4. Reflective practice;
5. Resources;
6. Soapboxing;
7. Other.

The next thing is to label a number of posts from the pilot sample with the appropriate category number. This was carried out be the researcher drawing on examples from the literature, and existing domain knowledge, to identify examples. Examples are given below. Note that the punctuation is missing as these examples are drawn from the training set which had already had it removed.

Category 1: CPD/Training/Advice
Blog posts in this category are about matters of Continuous Professional Development (CPD) for teachers, and training. They may write about what happened in a training session or conference, or suggest events teachers may want to attend. Typical keywords include 'CPD' and 'training', not always. The following example is a good one of 'advice':

"If you are a new teacher reading this you fall into one of two camps: 1., You have had a lovely honeymoon period with all your classes...".

Others that focus on more formal events include:

"Together with colleagues from teacher education institutions across the country I attended a full day of workshops and briefings hosted by the SQA..." and "Today I led the [...] Mathshub Curriculum Development meeting... [...] My presentation is (hopefully) embedded below...".

The extract from this post shows CPD in use:

"I have attended CPD events and read articles and blog posts in an attempt to see what everyone else is planning and to make sure I am on the right track., The change is so fundamental to everything ..."

Category 2: Positioning
'Positioning' is a blog post that expresses a belief or method that the blogger holds to be valid above others, and expresses this as a series of statements. Evidence to support the statements will usually be present, generally in the form of books or published research by educational theorists or other leading experts in the field of education. Examples include:

"Excuses disempower., Taking responsibility empowers., At Michaela we have a no excuses culture., What does this mean It means..."

"How do we measure the impact of technology used to support learning In the first of three posts I reflect on technology's role in education..."

The first example in particular is a good one for a series of firm, authoritative statements that were made by the Head of Michaela Community School. The second is chosen because the blogger has already flagged that they are going to write a series of posts, indication that they have planned their writing.

Category 3: Professional Concern
'Professional concern', then, is a also post expressing a view or concern, but the language will be more measured. Perhaps evidence from research or other bloggers will be cited, and the post will generally be longer. The blogger may identify themselves as a teacher of some experience, or perhaps a head of department or other school leader.

" Closing the gap: a well intentioned policy but one that worries me., I worry about its effect on our Aliyahs; that it encourages us to rein in their potential artificially..."

"Lesson observations: Approach with caution! [...] For example the MET project which spent millions of dollars..."

"Only a few days ago the Commission for assessment without levels published its final report., I have written recently about assessment without levels..."

The main emotion expressed in these posts is an element of 'concern' and 'worry' about changes in Education, and the impact this may have on students and staff.

Category 4: Reflective Practice
In this category, educators discuss their personal experiences of introducing a particular teaching strategy or approach in their own classroom.

Here is an example of a maths writing about their experience of teaching students how to work out the area of a figure:

"They know how to find the area of a rectangle., Do they though Give them an unlabelled rectangle and what do they do...".

Here, a second example makes clear that the teacher is blogging about something s/he has tried out in the classroom:

"I have been trying out some new assessment methods with my classes in the last couple of terms., This is partly a result of changes in the way I have taught some things..."

Category 5: Resources
As well as sharing resources directly, or providing URL links to the source, many posts in this category also offer ideas for ways to approach teaching in the classroom. This is not the same as 'reflective practice' - there is no attempt to *evaluate* the success of

the suggestions, but as every teacher knows a different route in to teaching a familiar subject can be really useful.

“I thought I would post the first three weeks of these here in case the questions were of use to anyone., Feel free to use as you wish....”

“Google Classroom Google’s tool for managing sharing and collaboration in your classroom has had an important update...”

This blog is a series of links to useful ideas and examples of visual displays:

“Hoping to add science careers Aegilopoides a display about women in science., Another pinterest style board Ljrn43 poems about the elements written by students...”.

Category 6: Soapboxing

‘Soapboxing’ is characterised feelings of anger expressed by the blogger. There may be a call to action, and the blog may be written as if the blogger were delivering a speech, as this example illustrates:

“Posters lessons., We have all done them., Maybe we were naive NQTs who believed it really would be a great way for 7 set 4 to synthesise their understanding of the formation of ox-bow lakes., Maybe we were exhausted the week before Christmas sleep-deprived to the point of torture and thought it would tide us over., Maybe it was coming up to the summer holidays...”.

The repetition of ‘maybe’ is a rhetorical device most often seen in speeches designed to persuade the audience to a particular viewpoint.

The other device used is the rhetorical question, for example:

“ Does she not get set homework Does she never get detentions Course not...”.

Other blog posts contain a clear sense of frustration:

“I sighed once again when I opened my laptop this morning and found this new article from secret teacher’ (a blog from various teachers wanting to bemoan the state of UK education each week under a veil of anonymity)., This week it is about how maths is useless...”

Category 7: Other (Uncategorised) In general, this should be posts not concerned with Education, although in practice - and to an algorithm - identifying posts in this way is probably challenging. This is discussed in more detail in chapter 6.

The labelled posts are used in step 6 to test a range of classifiers, before the most accurate classifier is deployed on the final set of data.

3.5.3 Algorithm Testing

Machine learning can be used to process and analyse language based on semantics or lexography. However, this is generally a computationally expensive approach; a common and often equally effective approach is to start by counting the number of words used across an entire corpus, and in *each document*. This is also referred to as a 'count vector matrix' or CV. This can be extended by weighing the score for each word according to the number of documents that contain the term, and the number of documents in the entire corpus i.e. TF (term frequency) x IDF (inverse document frequency: TFIDF. The result is that the words used many times within a small number of documents have the highest score; lower when the term occurs fewer times in a document, or occurs in many documents; lowest when the term occurs in virtually all documents (Manning et al., 2008, Ch.6, p.119). Neither vector takes into account word order or meaning - each blog becomes a 'bag of words'.

Once a matrix of CV or TFIDF scores has been generated, a distance-based model can be applied to measure the *closeness* between text objects. Several models have been developed, but one of the most common is calculating the *cosine similarity* (Russell, 2014). This measures the *angle* between two documents (see figure 3.10). The more similar the document, the smaller the angle. This is on example of how a 'distance model' (referred to in Chapter 2, section 2.3.2.1) is generated.

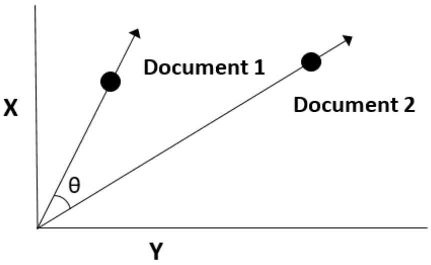


Figure 3.10: Cosine similarity

Once we have a vector space model, it is then possible to calculate where documents are in the space, and how near (or far) they are to all the other documents. From here it is relatively easy to cluster or classify documents into groups according to cosine similarity. This distance model (or models - CV and TDIF scores were tested) is used with the classifier and the clustering algorithm.

3.5.4 Classification Using Semi-Supervised Learning

Once the blog text data has been transformed and distance models generated, the process of classifying the corpus can begin. A small training set of blog posts were labelled according to the category definitions above, and used to train and test a range of classifiers. The choice of classifiers and results are discussed in the following Chapter 5.

3.5.5 Clustering Using Topic Modelling

Again, the blog text data is transformed and distance models generated. One of the most commonly used algorithms for topic modelling is Latent Dirichlet Allocation (LDA). LDA is an iterative algorithm with two main steps: at the initialisation stage, each word is assigned to a random topic (the number of topics having been declared in advance); the algorithm then iterates through each word, taking into consideration the probability of the word belonging to a topic, and the probability of the document to be generated by a topic. The result is a list of topics together with the frequency with which words appear in the topic in rank order from most to least. As before, the results of topic modelling were tested using a CV and TFIDF vectors.

However, while the number of categories used by a classifier is known in advance, this is not the case with topic modelling. The number of topics is set by the researcher, generally with little or no idea of the optimal number. It is entirely possible that no optimal number exists, and that if it were possible to map the distance model in 3D space, groups of data would not become apparent. In short, no *pattern* could be discerned by a human observer. There are methods of trying to ascertain the optimal number of topics, and these are discussed in more detail in Chapter 5. A pragmatic approach is to set the number of topics, and use an appropriate algorithm to compress and summarise the data into a 2D model, and view the results.

The following section will discuss the various ways the results of this research can be visualised (and how to judge the 'best' number of topics for topic modelling). The third research question concerning events in education such as policy changes or speeches and comments from ministers that impact the Edu-community is also best presented as a visual timeline. Therefore, the construction of timelines and data visualisation is presented in part 3.

3.6 Part 3: Building a Timeline and Data Visualisation

Building a timeline of events in Education is not as straightforward as it may seem. To begin with, there are Acts and White papers, committee reports, announcements and speeches; these can originate from the Department for Education as well as Secretary of State, or the Chef Inspector of Ofsted. However, a website charting the history of Education was used[11] as well as the 'News and Communication' section of the governments website gov.org with respect to Michael Gove.

An interactive version of the timeline based on events extracted from the Education England website was created using the online platform TimeLineCurator[12] and is available to view by following the link in the footnote. The categories on the interactive timeline are limited by the software used to create the visualisation. A condensed version from 2004 until 2017 is represented in Figure 4.5 in the following Chapter. An interactive timeline of Michael Gove's communications from 2010 to 2014 is available here[13](see footnote link).

The final step is to produce visualisations of the results of applying the classifier and topic modelling.

3.6.1 Data Visualisation: Classification Results

Visualising the results of a classifier is relatively straightforward, as each blog post in each year is labelled with a number representing a category label. Therefore, the number of entries in each category can be counted and represented as a series of bar charts. However, one of the most effective methods is to create a 'sankey' diagram, as shown in Figure 4.1 in the following Chapter. This was achieved using a web site `http://sankeymatic.com/` to generate the diagram.

3.6.2 Data Visualisation: Topic Modelling

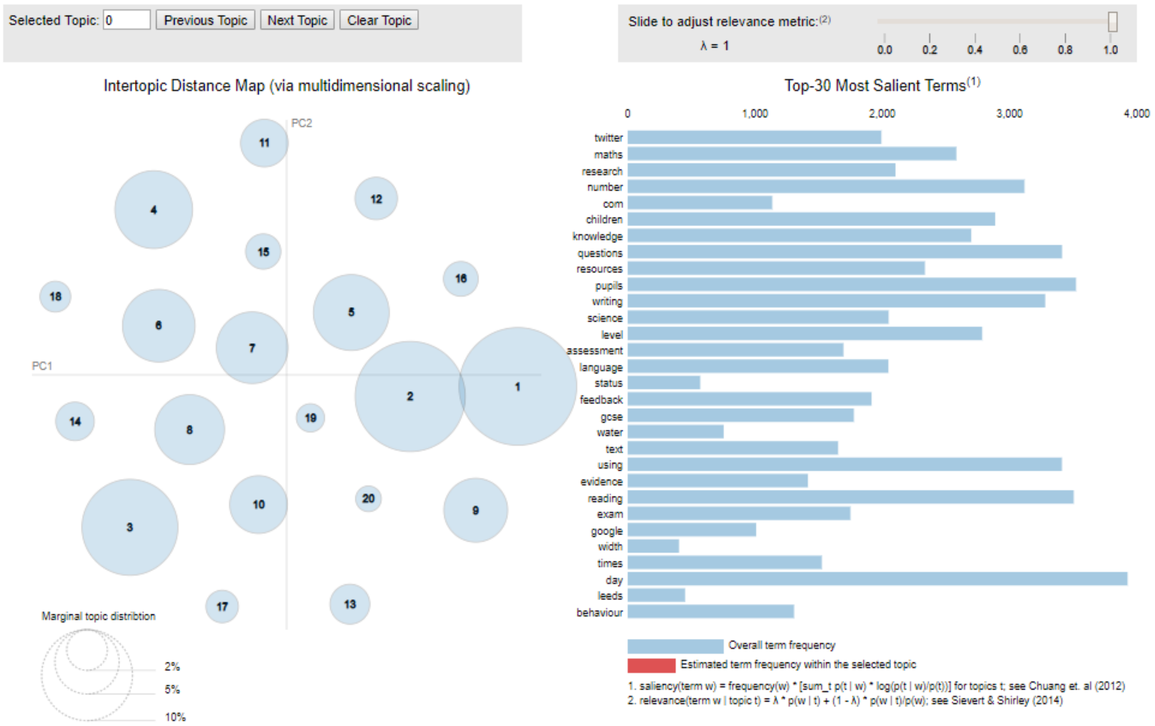
Data visualisation is more challenging with a clustering algorithm such as topic modelling. The data, which exists as a distance model in order to calculate which documents are most similar to one another, must now be summarised and represented in a two-dimensional space. Topic modelling also generates the top key words for each topic (which in turn are used to create a label for each topic) and these can also be added to a visualisation.

[11] http://www.educationengland.org.uk/history/

[12] http://www.cs.ubc.ca/group/infovis/software/TimeLineCurator/tlcExport/?tl=AnEducationTimeline

[13] http://www.cs.ubc.ca/group/infovis/software/TimeLineCurator/tlcExport/?tl=TheGoveYears20102014

One popular visualisation is shown in 3.11 (Sievert and Shirley, 2014). The circles represent a cluster of topics and the larger the circle, the more significant the topic within the corpus. To the right is a list of the thirty most significant terms across the whole corpus, and above this a slider that, when viewed as an interactive file allows more weight to be placed on the ratio of the frequency for a specific topic compared with the overall frequency of the word. There is an example made from the sample data here https://sh9g14.github.io/model.html.



Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3.11: Topic Modelling with Scikit-learn (*topics*=20)

We can observe from Figure 3.11 that the topics are well-separated in the space. More topics may result in the circles overlapping one another, less topics larger gaps between the spaces.

It is also possible to plot clusters of documents - data - using a more conventional graph where a coloured dot represents the document in a cluster, with all other documents in the grouo the same colour. However, where to draw the boundaries that separate the clusters will vary with the algorithm chosen to cluster the data. Where data has a non-linear structure, i.e. it cannot be divided in a meaningful way using straight lines, a clustering algorithm that can take better account of the structure might be more successful. One of these is a t-distributed stochastic embedding (t-SNE)(van der Maaten and Hinton, 2008) graph: a low-dimensional map which captures a non-linear structure. This is shown in Figure 3.12. We can observe that although the colour separation isn't perfect (each colour represents a topic), the topics do separate out reasonably well.

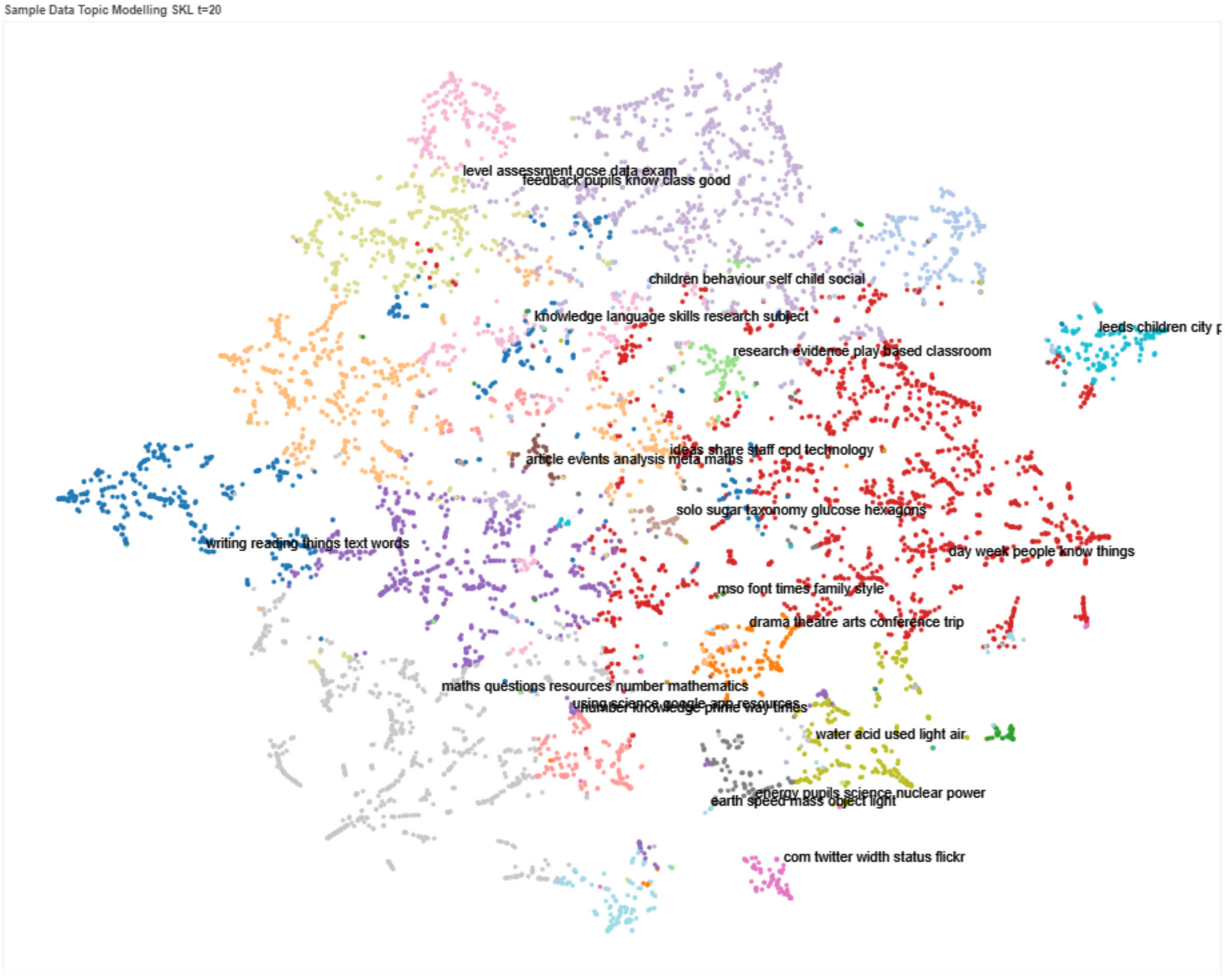


Figure 3.12: t-SNE representation of 20 topics generated by LDA on sample data

3.6.3 Topic Labelling

While the category labels are known in advance, the topic labels likely to arise from applying LDA are not. However, it is anticipated that the labels can be ascertained using the top 5 terms extracted from each cluster of posts. Some of the labels may be subject-specific i.e. ‘geography’; others may be broader such as ‘primary’. The number of clusters is also expected to vary year-on-year. Prior knowledge of the domain should be especially helpful here, as some of the references my be unknown to someone who has not worked in education e.g. ‘solo’ as part of ‘SOLO taxonomy’, a method of assessing student progress.

Furthermore, there are going to be many more topic labels than categories. The sample data on which the Figures above are based already indicate 20 topics; whiles some of these topics may persist across several years, new ones will arise. Given that Research Question 3 is looking for a link between Education policy and the topics discussed in the Edu-blogosphere, it is expected that at least some of the topic labels emerging from the data can also be applied to events from the timeline and announcements from the Department for Education. However, the labels will be generated from the results of topic modelling, and it is only after this has been done that the events from the timeline and the government communications will also be labelled. New labels may be needed.

3.6.4 Bubble Charts

A bubble chart is a type of scatter chart where the data points - topics - are represented as solid circular shapes, or bubbles. The topics would be represented in rows, the years in which they appear the columns. However, the bubbles have the added dimension of size which is directly linked to the *proportion* of a topic that is represented in the year. The larger the bubble, the more significant the topic. Thus, it is easy to get a good idea very quickly of both the *range* of topics discussed, and the *importance* of topics over time when one looks at a visualisation. The bubble charts created and presented in the next Chapter were created using Tableau[14].

3.7 Summary

The Methodology section of this research has been split across three sections. The way the raw data is treated *before* it can be analysed is important, particularly with regard to the stopwords removed before content analysis, which was discussed in part 2. Finally, this shorter section has provided methods for labelling and visualising the data. In the case of classification, this is a simple counting exercise. However, topic modelling presents different challenges as the shape of the underlying data and results are probalistic. Here, data visualisation has the additional function of being able to show if the pre-selected number of topics is reasonable or not.

The following Chapter will present the results of the two experimental approaches, classification and topic modelling, illustrated by the visualisations mentioned above.

[14] https://www.tableau.com/en-gb

Chapter 4

Results

4.1 Introduction

As we have seen in Chapter 1, the blogosphere is used by Edu-professionals in the UK to share ideas and resources, talk about their practice among themselves, and to express their opinions with regard to some the policy decisions made by the Department of Education which impact their work. As some blog posts appear to have been read by representatives of the current government, the Edu-blogosphere - or at least some members of it - can be said to have assumed a new importance as their voices have been heard. Whether they have been acted upon is a question this research does not address, but what it *does* seek to do is carry out a survey of as many Edu-blogpsphere posts as possible establish what topics have occupied it. The next step is to see if there have been any changes in focus year-on-year, and see if there is any evidence to link them to changes in Education policy.

A list of blogger URLs was used to extract the content of a large number of Edu-bloggers, and the data cleaned and processed in preparation for answering the Research Questions (detailed in Chapter 3). In accordance with the experimentation carried out on the parameters of the classifier and clustering algorithms, the data was:

- labelled with a category label using a small training set of pre-labelled data;
- assigned a dominant topic where the total number of topics to be assigned has been specified in advance.

The purpose of this chapter is to present the results using a range of data visualisations. A full discussion of the results will follow in Chapter 5.

Table 4.1: Research Questions

Research Question 1: What kind of subject matter, themes or issues have been discussed by the Edu-community since the arrival of the first available Edu-blog?

Research Question 2: Has the subject matter, themes or issues discussed by the Edu-Community changed since the arrival of the first available Edu-blog?

Research Question 3: Is there any link between the subjects, themes or issues discussed by the Edu-community and changes in Education policy?

4.2 Classification Using Semi-Supervised Learning

As stated previously, the use of a classifier that is built using categories drawn from existing research will not reveal anything new with regard to answering research question 1, but it may prove useful for research questions 2 and 3, and provide some oversight of the Edu-blogosphere in general. The classification model draws categories from the previous research into the content of blog posts from the Edu-community to define six categories, with a seventh being added to catch 'everything else' - and build a classifier. Each blog post is assigned to one of seven exclusive categories. The chosen model, a multinomial Naive Bayes classifier (the range of classifiers tested is discussed in detail in Chapter 5), reports an accuracy score of 62%. In short, *what* the Edu-community blogs about has already been answered by a review of the existing literature. We know from the research reviewed in Chapter 2 that teachers and other educators talk about the things that concern them during the course of executing their professional responsibilities: how they teach, the resources they have found useful, and the things that concern them about the directives they are given. Applying a classifier does not tell us anything new, although the distribution of the topics over time is more relevant for Research Question 2.

Figure 4.1 starts with the total number of blogs each year in the left column. The ordering of the years is an anomaly of the software used to build the model. It provides a useful visual summary of how the total number of blogs each year separate out into the different categories. The largest category overall is 'CPD/Training' (Continuous Professional Development), followed by 'Positioning'.

The second Figure 4.2 presents the same data, although this time the years are in date order. The colour bands across the columns represent the percentage each category makes up of the total number of blogs.

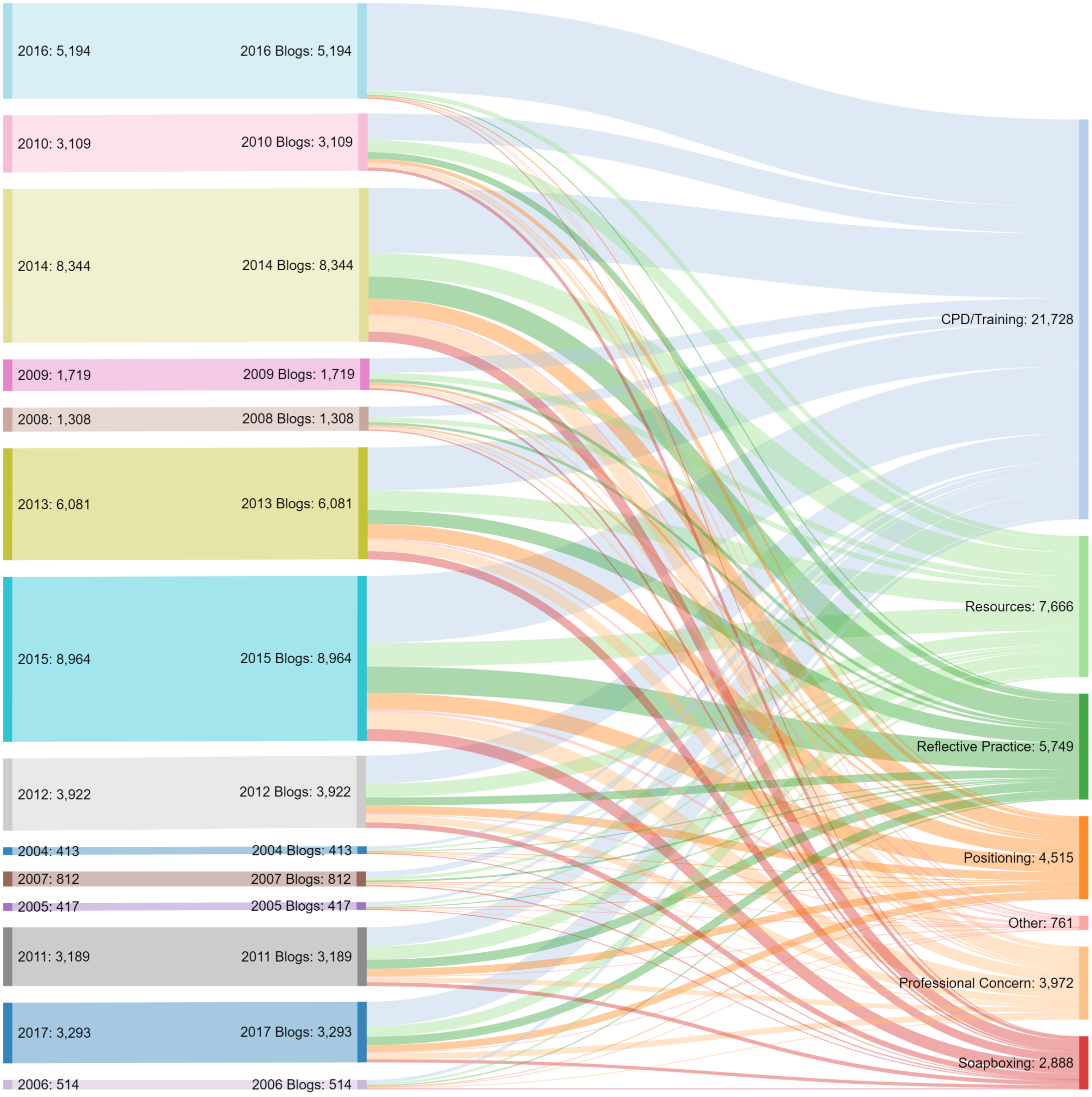


Figure 4.1: Total no. of blogs per year, split into categories

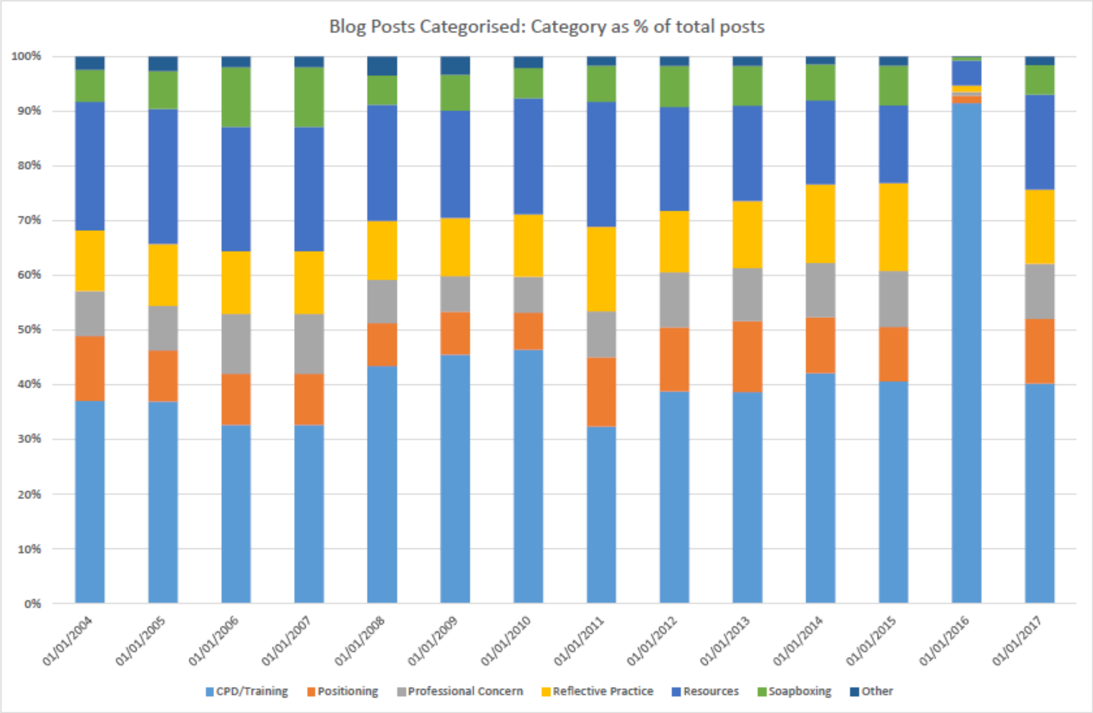


Figure 4.2: Categories as a percentage of the total no. of blogs

4.3 Clustering Using Topic Modelling

The second approach to answering the first research question was to apply a clustering algorithm to explore the blog posts with no *a priori* assumptions. Topic modelling using Latent Dirichlet Allocation (LDA) assumes that each document in the corpus contains one or more n-topics. Each topic is expressed as a probability within the document i.e document 1 may have a 60% probability of belonging to topic 1, 40% for topic 2 etc. In other words, a document - blog post - may 'belong' to one or more topics. Categories are not exclusive. As expected, this produced a much finer-grained result which suggested there were 53 topics within the entire corpus. A set of tables documenting, for each year of blog data, the following items is included in Appendix C:

- Topic number i.e. 1-15;
- number of documents (blogs) in each topic;
- topic labels (constructed manually);
- top 10 key words;
- final topic label.

Each of the topic categories is shown in figure 4.3. Here, the years are still shown in columns, with the topics in rows. A shortened topic key is shown at the left of the figure, with a more detailed list of the topics shown in table 4.2. The size of the bubble is in accordance with the number of blogs about the topic as a percentage of the total number of blogs that year; for example 'Assessment' (ASS) represented a significant proportion of the total number of blogs in 2013, 2014, 2015 and 2017.

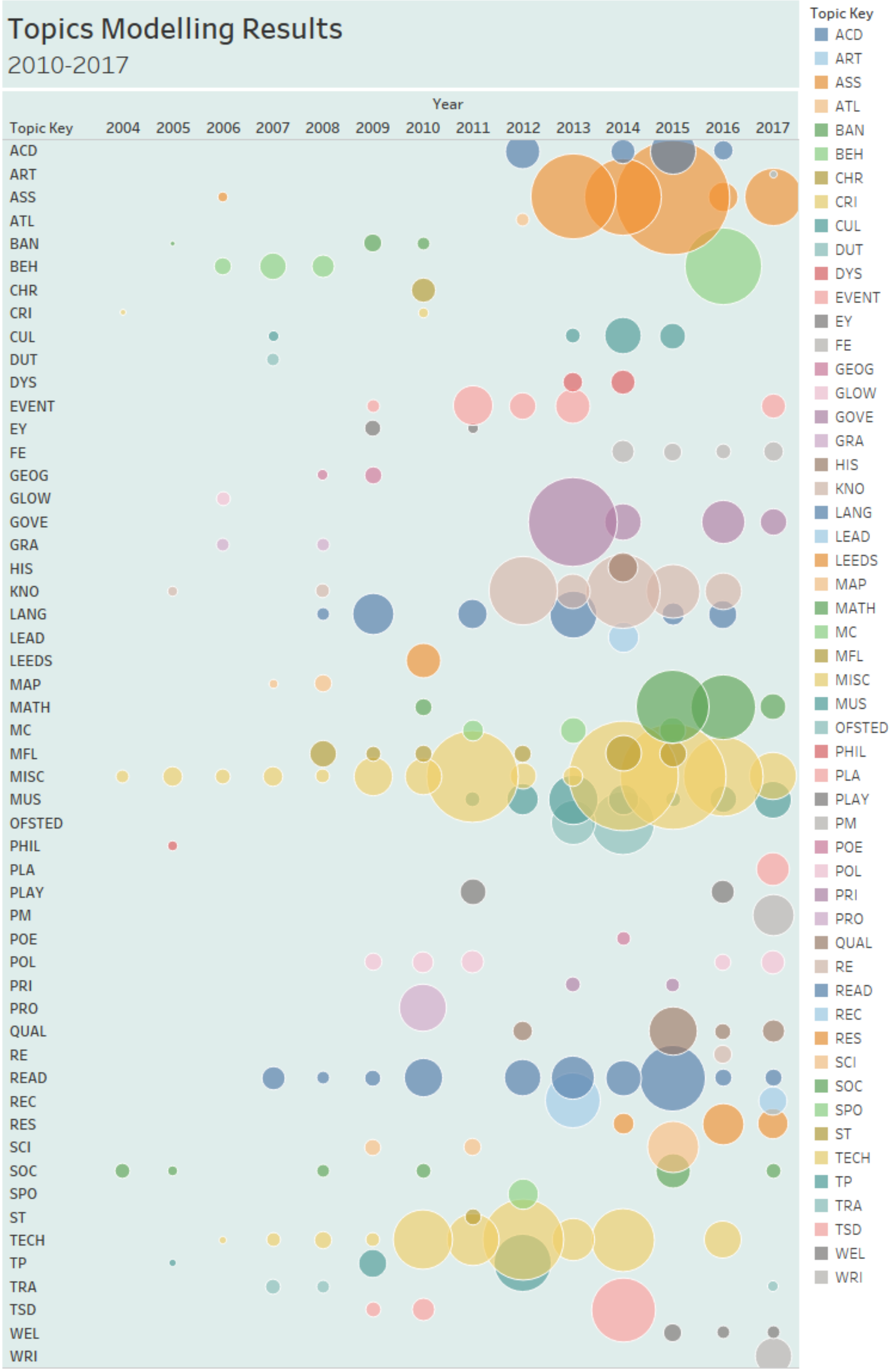


Figure 4.3: Topic Categories 2004 - 2017

TopicKey	Expanded	MainTopic	Expanded
ACD	Academy & Free Schools	DOE	Dept. for Education, policies
ART	Art	SUB	Subjects
ASS	Assessment & Feedback	TPR	Teaching practice
ATL	Association of Tchrs & Lecturers	STA	Staff issues, training
BAN	Banking	OTH	Miscellaneous
BEH	Behaviour	BHR	Behaviour, school policies
CHR	Christmas	SUB	Subjects
CRI	Crime	OTH	Miscellaneous
CUL	Culture	OTH	Miscellaneous
DUT	Duty	STA	Staff issues, training
DYS	Dyslexia, phonics	LIT	Literacy
EVENT	Conference	COM	Skills, knowledge, cognition
EY	Early years / nursery	CUR	Curriculum
FE	Further / higher education	CUR	Curriculum
GEOG	Geography	SUB	Subjects
GLOW	Glow Scotland	COM	Community, meeting, social
GOVE	Policy, Gove	DOE	Dept. for Education, policies
GRAM	Grammar schools	DOE	Dept. for Education, policies
HIS	History	SUB	Subjects
KNO	Knowledge (generally)	SKC	Skills, knowledge, cognition
LANG	Language, general	LIT	Literacy
LEAD	Leadership	STA	Staff issues, training
LEEDS	Leeds	OTH	Miscellaneous
MAP	Maps	OTH	Miscellaneous
MATH	Maths	SUB	Subjects
MC	Memory & Cognition	SKC	Skills, knowledge, cognition
MISC	Miscellaneous	OTH	Miscellaneous
MFL	Modern Foreign Languages	SUB	Subjects
MUS	Music	SUB	Subjects
OFSTED	Ofsted	DOE	Department for Education, policies
PHI	Philosophy	OTH	Miscellaneous
PLA	Play	OTH	Miscellaneous
PLAY	Plays e.g. Shakespeare	LIT	Literacy
PM	Planning & marking	TPR	Teaching practice
POE	Poetry (inc. specific poems)	LIT	Literacy
POL	Politics (general)	OTH	Miscellaneous
PRI	Primary	CUR	Curriculum
PRO	Project	OTH	Miscellaneous
QUAL	Qualifications / Exam boards	CUR	Curriculum
RE	Religions education, faith	SUB	Subjects
READ	Reading, books etc.	LIT	Literacy
REC	Recources	SUB	Subjects
RES	Research	TPR	Teaching practice
SCI	Science	SUB	Subjects
SOC	Social, networking	COM	Community, meeting, social
SPO	Sport	SUB	Subjects
ST	Solo Taxonomy	SKC	Skills, knowledge, cognition
TECH	Technology, ICT	SUB	Subjects
TP	Teaching practice	STA	Staff issues, training
TRA	Travel	OTH	Miscellaneous
TSD	Training & staff development	STA	Staff issues, training
WEL	Welsh (language)	OTH	Miscellaneous
WRI	Writing	LIT	Literacy

Table 4.2: Topic Modelling Categories

In summary, a classifier needs to have a set of labelled data to work with. This labelled data is derived from the research literature, and therefore we can conclude that Research Question 1 has already been answered - it is the distribution of the data that is important, and more applicable to answering Research Questions 2 and 3. Topic modelling provides a completely fresh response to the themes and issues discussed in the Edu-blogosphere, and indicates that the topics are numerous.

4.4 Topic Labelling

For the classification model, it was possible to represent all seven categories, and their changing distribution over the course of fourteen years (see Figures 4.1 and 4.2). There has been some change over time, but it is limited and generally the balance between categories remains consistent. Possible reasons for this are discussed in the following chapter.

Topic modelling produces a more complex model, with over 50 topics labelled. Figure 4.3 indicated clearly through the size of the topic bubbles that there *are* changes over time, but in order to see this more clearly the topics have been condensed down to a total of ten key or 'main' topics, shown in table 4.2 as the two columns at the far right of the table. The list of summarised topics is as follows:

- BHR: Behaviour;
- COM: Community, meeting, social, networking, conferences);
- CUR: Topics to do with the curriculum (Including FE and HE);
- DOE: Department for Education matters, policies etc.;
- LIT: Reading, writing and matters to do with literacy;
- SKC: Skills, knowledge, cognition;
- STA: Staff issues, training;
- SUB: Subjects (may be specific such as MFL (modern foreign languages) or inferred);
- TPR: Teaching practice;
- OTH: Miscallaneous items (either not related to teaching, or unclear from the topic key words);

A visual representation of the main topics only is shown in Figure 4.4. Note that the category 'Other' has been removed. As in the first graph, the size of the bubbles represents the proportion of the blogs in the category compared with the others i.e. the category with the highest number of blogs for that year will have the largest bubble.

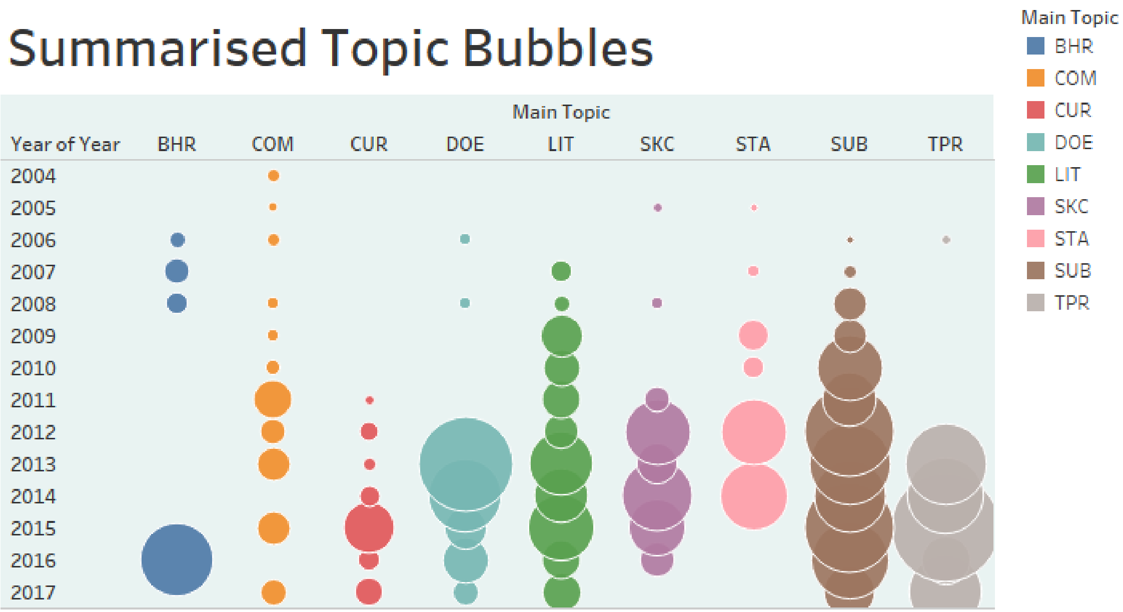


Figure 4.4: Main Topic Categories 2004 - 2017

4.5 The Relationship Between The Results

Two distinct approaches to exploring and analysing the Edu-blogpsphere have been used. There appears to be little, if any, correlation between the two and this is certainly the case in terms of the way the different algorithms work. A classifier has labels assigned to each category from the outset; topic modelling requires that they be devised once the results are known. Furthermore, the categories that emerged from the literature were extracted from usually small sets of data, and were looking for themes in terms of the matters that would concern teaching *generally*, rather than subject-specific matters. This is discussed in Chapter 2, Section 2.2.1. The two reviews of online teaching communities (Macia and Garcia, 2016; Lantz-Andersson et al., 2018) do not include any studies carried out using computational methods from Computer Science (with the exception of social network analysis). This research analyses blog posts *at scale* and over time, and could not be achieved using hand-coding.

As mentioned above, a classifier does not reveal any new content, only the *distribution* of the content over time. The Naive Bayes classifier such as the one used here uses *conditional probability* to classify blogs i.e. given the content as a labelled post, an unseen post with similar content is likely to be in the same category. The probabilty increases with similarity to the labelled set of blogs. Topic modelling performs in a similar way but without the benefit of labelled posts. However, this has resulted in a much finer-grained range of topics that are of more interest to the community. This is discussed in more detail in the following chapter.

In short, each analysis of the year-by-year sets of Edu-blogs is unconnected with the other. It is certainly true that any of the categories may also contain topics from topic modelling - i.e. ‘Reflective Practice’ almost certainly includes blogs about ‘maths’, ‘science’ or even ‘Ofsted’ but there is no correlation in the statistical sense between the two. What they offer are *different* ways of examining the same data, offering different insights.

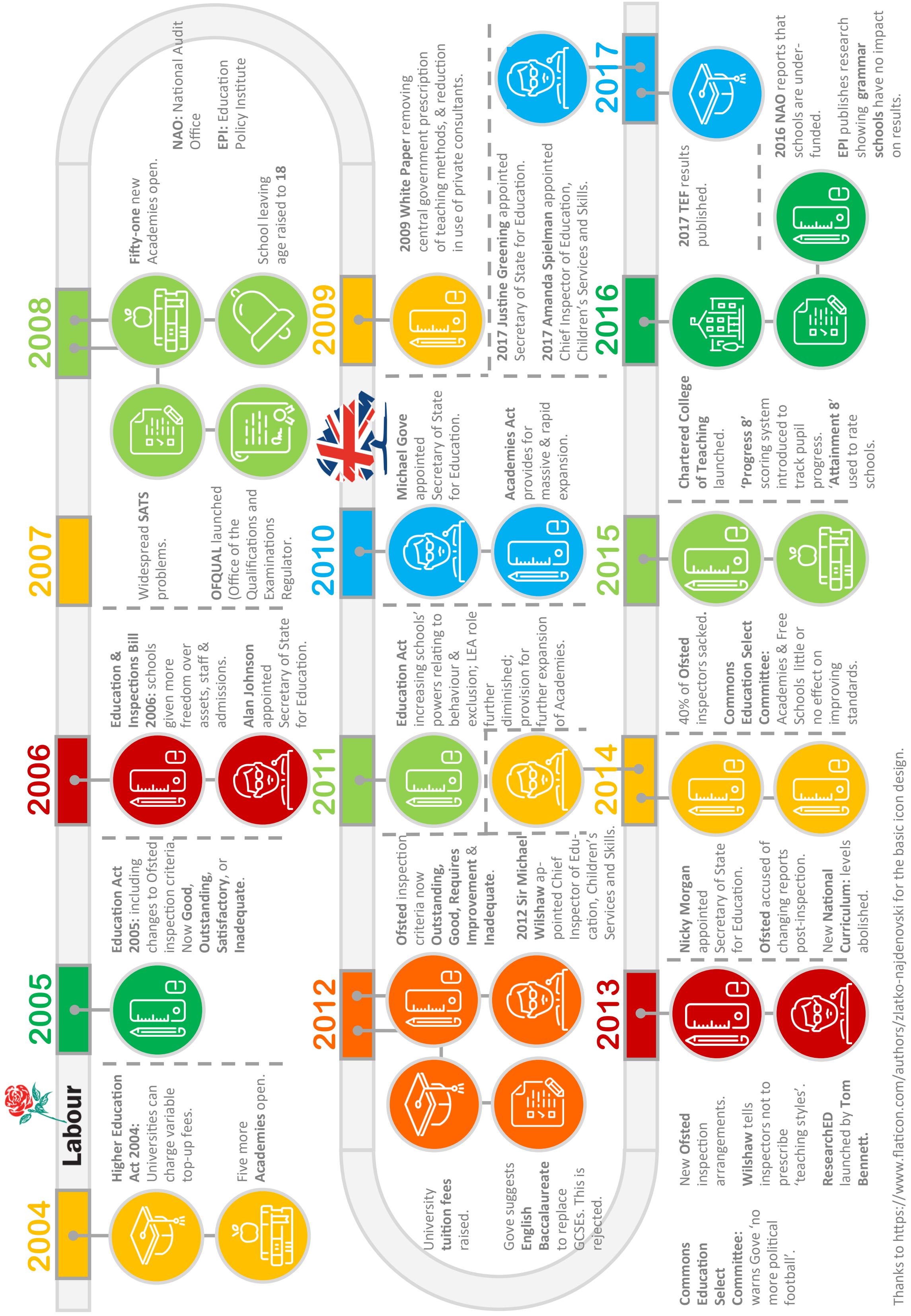
4.6 Building A Timeline of Events in Education

However, before examining these results in detail, Research Question 3 asks if changes in the Edu-blogosphere can be linked with changes in Education policy, including the actions of the schools’ inspection body, Ofsted. In order to do this, it is necessary to create a timeline of these events, which is presented in Figure 4.5.

The event which first brought the Edu-blogosphere to public attention started with the appointment of Michael Gove as Secretary of State for Education in June 2010. The Conservatives’ commitment to increasing the number of academies following the Academies Act was evident by February 2014 when 3,657 Academies had been opened[1]. Other events included the raising of tuition fees, and the abolition of levels (used to set a benchmark for where children were ‘at’ following their SATS in primary school, and to set targets for achievement at GCSE). Ofsted inspection criteria were altered in 2012, dropping the ‘satisfactory’ judgement and replacing it with ‘requires improvement’ (RI). A school judged to be ‘requiring improvement’ was expected to convert to an Academy, usually coupled with a change of Head teacher and senior staff.

The full list of the policies, events, announcements and personnel changes are shown in Figure 4.5 and Figure 4.6. As far as possible, the colours in Figure 4.6 are consistent with the ‘Topic Modelling Results’ in Figure 4.3.

[1] https://web.archive.org/web/20140223002024/http://www.education.gov.uk/schools/leadership/typesofschools/acade
academies



Higher Education Act 2004: Universities can charge variable top-up fees.

Five more **Academies** open.

Education Act 2005: including changes to Ofsted inspection criteria. Now **Good, Outstanding, Satisfactory,** or **Inadequate**.

Education & Inspections Bill 2006: schools given more freedom over assets, staff & admissions.

Alan Johnson appointed Secretary of State for Education.

Widespread **SATS** problems.

OFQUAL launched (Office of the Qualifications and Examinations Regulator).

Fifty-one new Academies open.

School leaving age raised to **18**

NAO: National Audit Office

EPI: Education Policy Institute

2009 White Paper removing central government prescription of teaching methods, & reduction in use of private consultants.

2017 Justine Greening appointed Secretary of State for Education.

2017 Amanda Spielman appointed Chief Inspector of Education, Children's Services and Skills.

Michael Gove appointed Secretary of State for Education.

Academies Act provides for massive & rapid expansion.

Education Act increasing schools' powers relating to behaviour & exclusion; LEA role further diminished; provision for further expansion of Academies.

Ofsted inspection criteria now **Outstanding, Good, Requires Improvement** & **Inadequate**.

2012 Sir Michael Wilshaw appointed Chief Inspector of Education, Children's Services and Skills.

University **tuition fees** raised.

Gove suggests **English Baccalaureate** to replace GCSEs. This is rejected.

Commons Education Select Committee: warns Gove 'no more political football'.

New **Ofsted** inspection arrangements.

Wilshaw tells inspectors not to prescribe 'teaching styles'.

ResearchED launched by **Tom Bennett.**

Nicky Morgan appointed Secretary of State for Education.

Ofsted accused of changing reports post-inspection.

New **National Curriculum:** levels abolished.

40% of **Ofsted** inspectors sacked.

Commons Education Select Committee: Academies & Free Schools little or no effect on improving standards.

Chartered College of Teaching launched.

'Progress 8' scoring system introduced to track pupil progress.

'Attainment 8' used to rate schools.

2017 TEF results published.

2016 NAO reports that schools are underfunded.

EPI publishes research showing **grammar schools** have no impact on results.

Thanks to https://www.flaticon.com/authors/zlatko-najdenovski for the basic icon design.

Figure 4.5: A Timeline of Events in Education

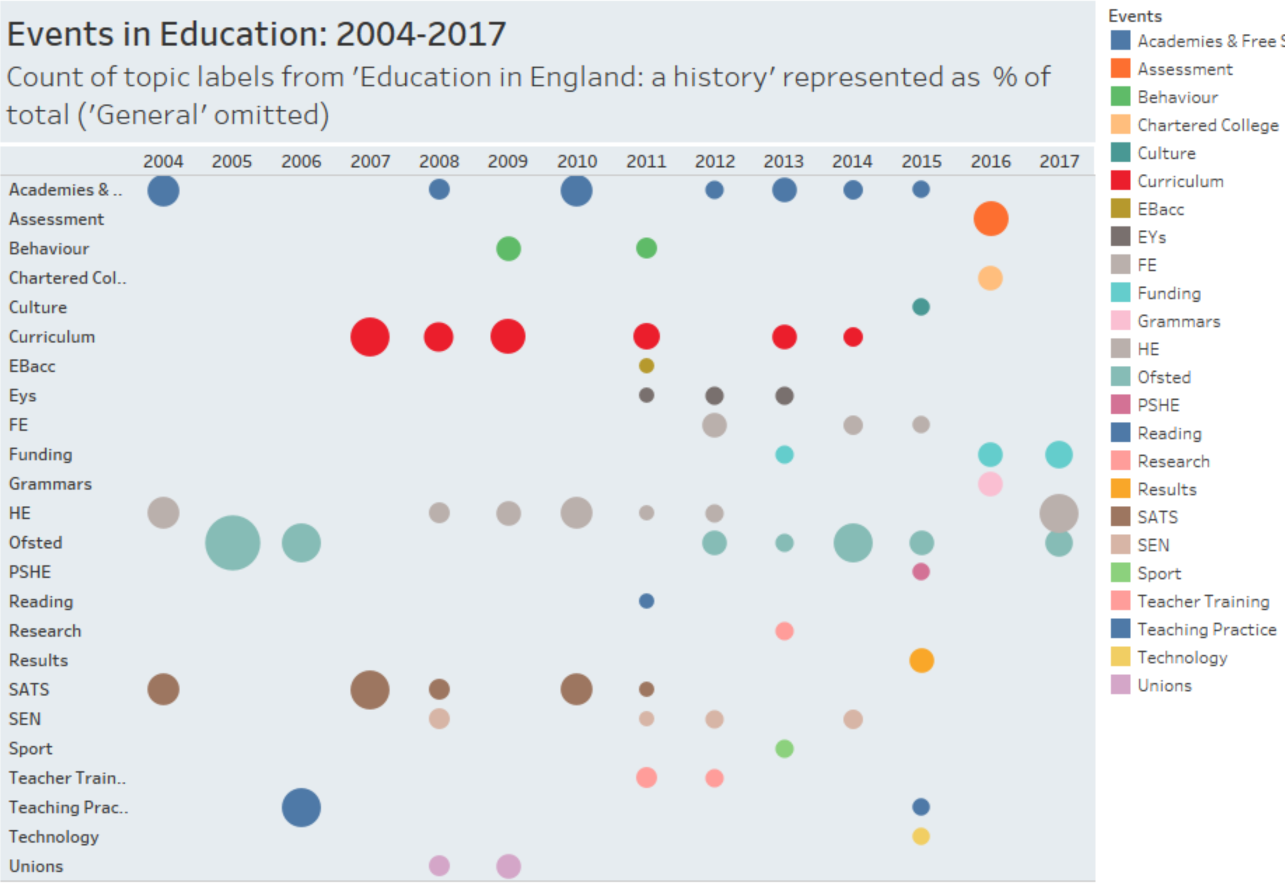


Figure 4.6: Events from the website 'Education in England: a history'

4.7 Conclusion

As stated at the beginning, the purpose of this chapter is to present the results with some brief additional information to provide context. The following chapter discusses the results in more detail, including offering some reasons behind some of the more interesting aspects of them.

An interactive representation of the Edu-blogosphere from 2004 to 2017 is available on this website https://sh9g14.github.io/testdata.html, with a corresponding bubble chart similar to Figure 4.4.

Chapter 5

Further Discussion of Results

5.1 Introduction

This chapter will look in more detail at some of the main points arising from this research. These will be divided into two sections: the transformation of the data, and the analysis of the content. This will include some of the limitations of the methods used, as well as the ways in which these have been addressed by the researcher. It is worth mentioning that, in light of the intention to bring the application of tools from Computer Science to answer research questions arising from Social Science, it has not been clear how to differentiate aspects of methodology from results and discussion as they each provide insight to the other. Therefore, there is some overlap.

The following sections will return to each of the 'three pillars' from Figure 1.1 and discuss some of the key findings in more detail.

5.2 Data Transformation

The quality of the results resulting from the application of any algorithm on the data are, to a large extent, dependent upon the quality of the data going in. One of the most significant early steps is the removal of stopwords. As explained in Chapter 3, section 3.4, this is an iterative process in which the set of stopwords provided by Scikit-learn was used, supplemented by some of the most frequently used words in the sample data set. Some additional words were also added as the process of testing the topic modelling algorithm proceeded. As discussed in Chapter 2, section 2.3.1, and investigated by (Nothman et al., 2018), there is no 'industry standard' list of stopwords, and any list used as a basis to begin removal should be adapted for the specific domain. In the process of building and testing a classifier, the algorithm used to build a matrix of word counts was adjusted to remove all words occurring in less that 1% of documents. The list of stopwords, the words included for the classifier and topic modelling, and the parameters of the topic modelling algorithm would all have had some impact on the final results. However, during the course of assessing optimum number of topics for each year - a process of entering different numbers, and judging the results visually and using the top 5 words per topic - the impact would not have been significant overall.

5.3 Word Frequency Counts

Even when the data has been cleaned and stopwords removed, there are still some steps to take before the final algorithms can be run on the data. A classifier and a clustering algorithm each rely on counting the words in each document, including words that appear in one blog post but not another i.e. zero counts. There are two methods of doing this, and each needs to be generated before they are tested with algorithms.

As discussed in Chapter 3, there are two matrices that can be passed to the classifier or the topic modelling algorithm: a ‘count vector’ (CV) matrix or a ‘term frequency, inverse document frequency’ (TFIDF) matrix. Using TFIDF, the less-frequently used words become more significant, and should represent the underlying topic with greater accuracy. It is also possible to adjust the parameters of the TFIDF and CV algorithms used by Scikit-learn to remove words that only occur in a *minimum* number or proportion of documents, or a *maximum* number or proportion of documents. In other words, it is possible to remove words that appear in *less than* x-number of documents, or in *more than* x-number of documents. This seems a sensible adjustment, given that there may be more value-less words in the corpus than just those included in the list of stop words; conversely, some words such as fragments of URLs embedded in the blog posts may represent a word used infrequently, and could therefore also be removed with no or little loss of valuable data. Both of these adjustments were used to pre-process the data prior to building the models used for categorisation.

The results of generating a CV or TFIDF matrix are used to create a vector space model, which is then used to build a *representation* of the data in two- or three-dimensional space. The underlying shape of this data can have a significant impact of the results obtained from a classifier and topic modelling. The results of testing both approaches with a classifier and topic modelling are presented below.

5.4 Selecting and Testing A Classifier

A *classifier* involves the supervised learning of a function that is nominal i.e. drawn from a finite set of possible values. In the case of this research, the problem is *multi-class* or *multi-nominal.* The classifier predicts the class of a proportion of the already labelled data and compares this with the actual class to produce an accuracy score. The 'trained' model is then deployed on the remaining unlabelled data. This is also described as a *probalistic* approach as the labelling of the previously unseen data is based on the *probability* that is belongs in class *a* and not class *b* or *c*.

The procedure, after preparing the data, is to split the data into a training set (75%) and a test set (25%), and then construct a TFIDF matrix and a word count matrix (CV). It is important to split *before* vectorizing because it is important to simulate 'the real world' where our future data may contain words not seen before. In this case, our 'future' data is our unlabelled data. Once a successful model has been established, this will be trained on the remaining data.

The following algorithms were trained on a sub-set of data drawn from the training set, and then tested on the remaining data.

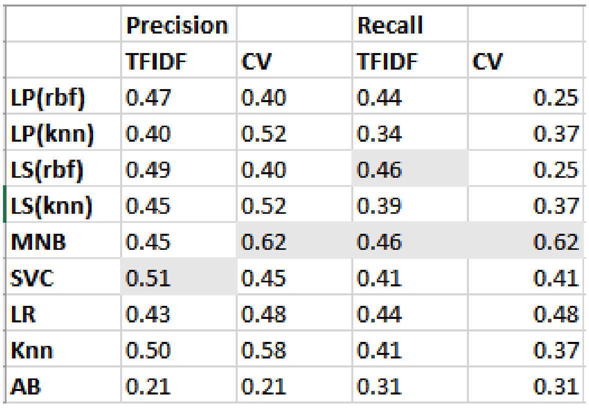
- Label propagation (rbf[1]);
- Label propagation (knn[2]);
- Label spreading (rbf);
- Label spreading (knn);
- Multinomial naive bayes;
- SVC (rbf);
- Logistic regression;
- Knn (K-nearest neighbours);
- Ada Boost.

Each algorithm approaches the task of classification slightly differently.

Label propagation, label spreading and ***K-nn*** all use a distance model to construct a feature space. The *K-nn* classifier assigns a label to each point in the feature space, which divides the space into regions within which points have the same label. The boundary separating these regions defines the unlabelled data that belongs to the

[1] 'rbf' (radial based function) and 'knn' (k-nearest neighbours) refer to the kernel method used by the algorithm. They calculate similarity based on distance in slightly different ways.

[2] see above.



	Precision		Recall	
	TFIDF	CV	TFIDF	CV
LP(rbf)	0.47	0.40	0.44	0.25
LP(knn)	0.40	0.52	0.34	0.37
LS(rbf)	0.49	0.40	0.46	0.25
LS(knn)	0.45	0.52	0.39	0.37
MNB	0.45	0.62	0.46	0.62
SVC	0.51	0.45	0.41	0.41
LR	0.43	0.48	0.44	0.48
Knn	0.50	0.58	0.41	0.37
AB	0.21	0.21	0.31	0.31

Figure 5.1: Classifier Scores using a range of Classifiers

group. The **Label propagation** and ***label spreading*** algorithms ‘spread’ the known labels through the nearest features until they meet a different labelled group. The **SVA** algorithm is a member of the class of algorithms known as SVMs (see Chapter 2, section 2.3.2.2), which ‘slice’ groups of features using the already-labelled vectors as ‘support’. The algorithm maps the features in a multi-dimensional space (based on a distance model) and looks for ‘valleys’ in the data (areas where data is sparsely clustered). **Logostic regression** works in a similar way, but the slice is *linear* and based on the *probability* that a given feature belongs in a specific class. **Multinomial naive Bayes** does *not* work with a distance model, but uses a word count matrix to calculate the probability that a feature indicates a vector (document) belongs in category a as opposed to any other category. **AdaBoost** is what is known as an ‘ensemble method’ as it combines a number of weak classifiers to create a strong classifier. The algorithm deployed within Scikit-learn “...begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases”[3].

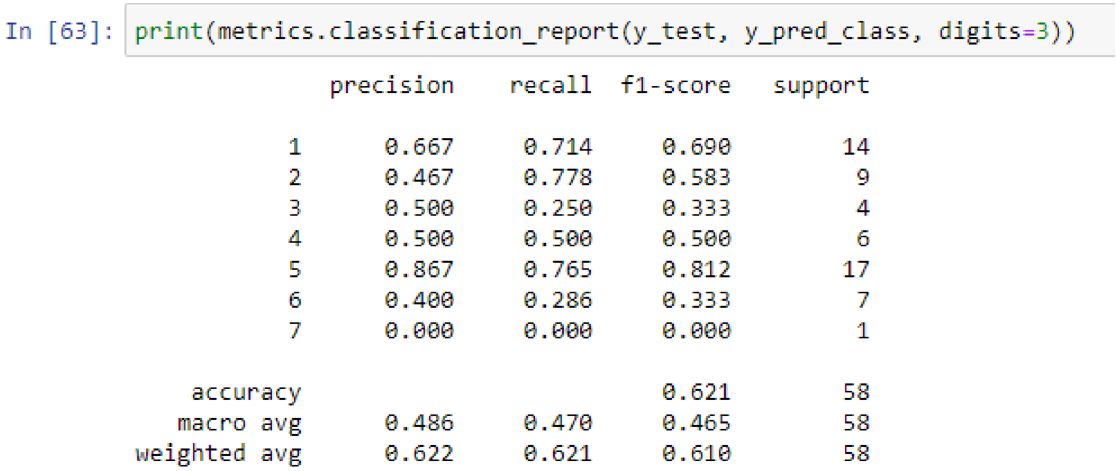
When evaluating the ‘success’ of a classifier, there are two results that need to be taken into consideration: precision and recall. ‘Precision’ is the number of correctly predicted posts in, say, class 1. For example, referring to figure 5.3, the algorithm multinomial naive Bayes predicted precisely 10 labels (0.667 or 10 out of 14, see figure 5.1) for class 1. ‘Recall’ is the correctly predicted number of posts in class 1 out of the number of *actual* class 1 labels; 0.714 in figure 5.1, or 10 out of 15. Which score is ‘best’ depends on the reason for calculating the scores - if trying to predict how many patients will develop diabetes, for example, the ‘recall‘ measure will be the most important, the reason being that it is better to predict that someone *will* develop diabetes even if they don’t (false negatives), than to just stick to the predicted positives. For other problems, *precision* may be more important. Figure 5.1 shows the precision and recall results for the classifiers listed above. The blog posts were either pre-processed using TFIDF or a

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

simple CV. According to figure 5.1, multinomial naive Bayes returns a reasonable score for precision and recall using a count vector (CV). Other algorithms such as Knn (see figure 5.1) vary in precision v recall, neither of which are as ‘good’ as multinomial naive Bayes (MNB). Pre-processing using TFIDF or a CV were evaluated. As figure 5.1 shows, the results were varied by multinomial naive Bayes generally returned the ‘best’ overall result.

It should be remembered that:

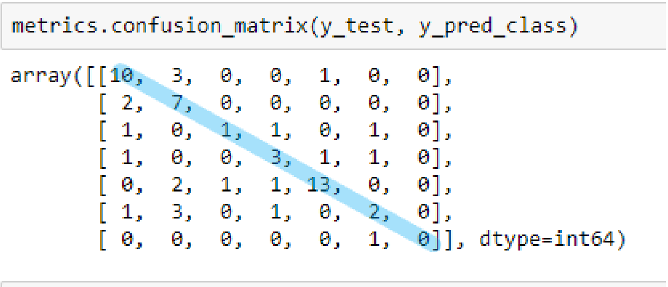
- the training set is relatively small (218 posts);
- the posts have been labelled by the researcher;
- ‘correct’ (the precision score) is correct according to class label of each blog post in the test set matching with the label that has been allocated in the training set.



```
In [63]: print(metrics.classification_report(y_test, y_pred_class, digits=3))

                 precision    recall  f1-score   support

            1        0.667     0.714     0.690        14
            2        0.467     0.778     0.583         9
            3        0.500     0.250     0.333         4
            4        0.500     0.500     0.500         6
            5        0.867     0.765     0.812        17
            6        0.400     0.286     0.333         7
            7        0.000     0.000     0.000         1

     accuracy                            0.621        58
    macro avg        0.486     0.470     0.465        58
 weighted avg        0.622     0.621     0.610        58
```

Figure 5.2: Metrics for multinomial naive Bayes

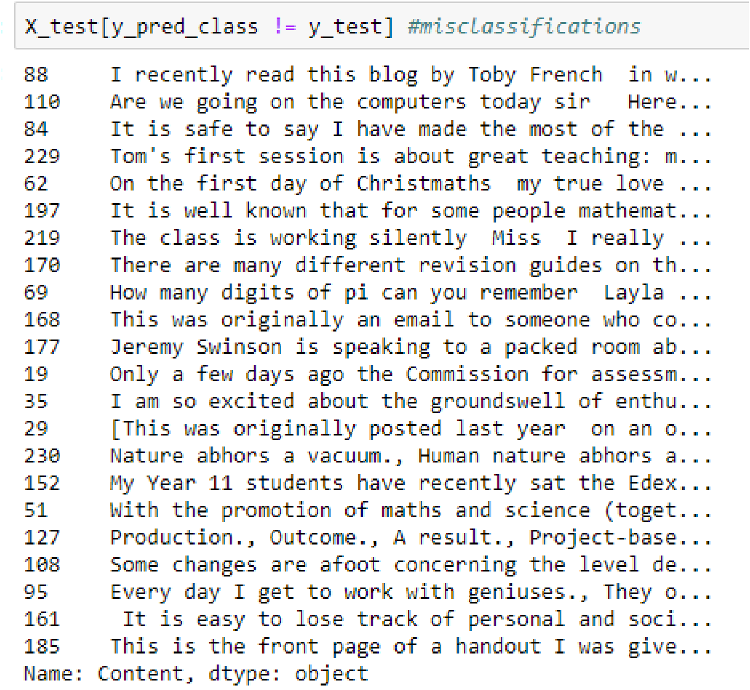
As stated above, the model that returns the ‘best’ accuracy score is multinomial Naive Bayes, using a count vector matrix. The confusion matrix (see figure 5.3) reveals the mis-classifications: each row and column represent categories 1-7; the correct predictions are represented by the numbers under the diagonal line. All other predictions are false.



```
metrics.confusion_matrix(y_test, y_pred_class)

array([[10,  3,  0,  0,  1,  0,  0],
       [ 2,  7,  0,  0,  0,  0,  0],
       [ 1,  0,  1,  1,  0,  1,  0],
       [ 1,  0,  0,  3,  1,  1,  0],
       [ 0,  2,  1,  1, 13,  0,  0],
       [ 1,  3,  0,  1,  0,  2,  0],
       [ 0,  0,  0,  0,  0,  1,  0]], dtype=int64)
```

Figure 5.3: mNB confusion Matrix

A closer look at the errors show which posts have been mis-classified (see figure 5.4). The temptation now would be to identify the *predicted* class and ascertain if the blog post might have been mis-classified in the first place. However, this is exactly the kind of approach that introduces bias into the model. Classifying blog posts into n-categories is difficult for a human, and it is entirely likely that even a group of humans with specialist knowledge of the domain i.e. teachers would not agree among themselves. The final step, therefore, is to save the multinominal Naive Bayes model, and apply it to the rest of the data.



```
X_test[y_pred_class != y_test] #misclassifications
```

```
88     I recently read this blog by Toby French  in w...
110    Are we going on the computers today sir   Here...
84     It is safe to say I have made the most of the ...
229    Tom's first session is about great teaching: m...
62     On the first day of Christmaths  my true love ...
197    It is well known that for some people mathemat...
219    The class is working silently  Miss  I really ...
170    There are many different revision guides on th...
69     How many digits of pi can you remember  Layla ...
168    This was originally an email to someone who co...
177    Jeremy Swinson is speaking to a packed room ab...
19     Only a few days ago the Commission for assessm...
35     I am so excited about the groundswell of enthu...
29     [This was originally posted last year  on an o...
230    Nature abhors a vacuum., Human nature abhors a...
152    My Year 11 students have recently sat the Edex...
51     With the promotion of maths and science (toget...
127    Production., Outcome., A result., Project-base...
108    Some changes are afoot concerning the level de...
95     Every day I get to work with geniuses., They o...
161     It is easy to lose track of personal and soci...
185    This is the front page of a handout I was give...
Name: Content, dtype: object
```

Figure 5.4: Mis-Classifications

5.5 Category Labels and Evaluation

A review of the existing literature concerning blogs written by Edu-professionals suggested six distinct categories. The categories were devised by the authors using qualitative techniques and often small sample sizes. What is noticeable is that the categories can be seen as 'themes', especially when compared with the results of topic modelling (discussed below). The qualifying blogs remaining in the data set (see Chapter 5) were allocated one of the seven categories based on the information learned by a classifier from a training set of labelled data: six categories from the literature review, and one final category 'other' for blogs that ostensibly were not concerned with Education.

Training a classifier on data, and using this on the remaining data to predict categories does produce consistent results i.e. the documents will always be allocated to the same category in every run, but the data going in will still obviously impact the results coming out. A range of classifiers were tested on the data after stopwords had been removed. The most accurate - multinomial naive Bayes (mNB) - returned an accuracy score of 62%. A better way of looking at this would be to state that, given a sample from a pre-labelled set of data, the algorithm was able to correctly predict the correct label of the remaining set of data i.e. the predicted label matched the *actual* label 60% of the time. This may seem low when we consider that 40% of the time the classification is wrong, but we must bear in mind that there are 7 categories. The probability of correctly predicting a category by chance is 17 or 14.29%, or an 86% chance of an incorrect prediction. It should also be remembered that a classifier will classify *every* blog in to one of the seven categories. It is impossible to say if the categories drawn from the literature is exhaustive. Furthermore, one class that remains problematic is that of 'other'. As a human, we might regard an 'other' or 'miscellaneous' category as 'everything that doesn't fit elsewhere'. A machine learning algorithm cannot make the same distinction - the mNB classifier has taken the examples of 'other' fed to it through the labelling of a training set and applied those rules to the unlabelled data. Therefore, blogs in this class are more likely to match the words used in the labelled training data than be a set formed of 'everything not in the other sets'. Nevertheless, this is probably the best solution when the focus is on the *labelled* categories and the classification of the information they contain.

The second thing to be considered is the classification of documents for the training set. This was carried out as part of the work of this thesis by myself. Inter-coder agreement between two or more coders classifying documents into several categories is challenging. An interesting example of this is a paper looking at the classification of a set of documents used in a wide variety of applications: the Comparative Manifesto Project (CMP)[4](Mikhaylov et al., 2012). Here, a set of party manifesto leaflets were coded by

[4] https://manifesto-project.wzb.eu/

a single coder against 57 possible coded categories (including 'uncoded'). The authors found that, when they repeated the coding experiment using 39 coders, two manifestos (and a further test combining both documents), they were only able to record reliability scores against the 'masters' of 0.43, 0.54 and 0.46 respectively. Krippendorff (2004) states that while there is no definitive answer to what represents a 'good' reliability score, that anything below .80 (and no lower than .667) should "...only be used for drawing tentative conclusions" (Krippendorff, 2004, Ch.11, p.241). The main points to draw from this is firstly that the mNB classifier returning an accuracy score of 0.62 based on nothing more than word counts is reasonable, and were we to use topic modelling to provide a basic framework to suggest topic categories for human coders, given that it returned a suggested 53 this would be a considerable challenge to inter-coder reliability. Secondly, this limitation was not addressed in this research because the aim was to provide a broad overview of the Edu-blogosphere; classification or clustering 'accuracy' is something future work may address by building on the work presented here. The challenge remains one of researcher bias - it is the researcher's own knowledge of the domain that is brought to the classification process, and only this is deployed to label a training set.

We are also assuming that the accuracy score translates to the unlabelled data in the same way. While this may have been demonstrated with other data sets, these sets are invariably either small, well cleaned and pre-processed, or both. They are often *already* grouped according to topic such as the 20 Newsgroups data set[5] and it is unsurprising that tutorials using them to demonstrate how classifiers or clustering algorithms work return accuracy scores of 90% or more. 'Real world' data is very different and there is therefore an element of trust that the unlabelled data is being allocated a label with the same degree of accuracy as the training set.

While the category headers were drawn from the existing literature, the interpretation of which blogs best fit into which category was a matter of making judgements based on the meaning of the language used. The difference 'Soapboxing' and 'Professional Concern' was also often a matter of tone as interpreted by the researcher. Blog posts are by their very nature more informal writings, and therefore some of the more formal language that would be present in, for example, research papers or articles written for journals are rarely present (as was often the case in the research of Murakami et al. (2017)). This often made it difficult to decide if a blog post was the more formal 'Positioning' or 'Professional Concern', when the language used was more appropriate for a general audience and presented in the diary-like format of a personal blog post. The nuances of meaning are of interest to cognitive scientists who are concerned with how humans learn language in the first place. This is discussed briefly in *Introduction to Semi-Supervised Learning* (Zhu and B, 2009, Ch.7), although the problem is classification of objects, not the subject of blog posts which is discerned though the understanding of language.

[5] http://qwone.com/ jason/20Newsgroups/

Furthermore, the problem in this research is one of classification into more than two categories, which makes the problem much harder.

In short, data cleaning, preprocessing, stopword removal, labelling a small training set by hand, calculating a distance model and creating a word count matrix are all examples of stages where the results can be impacted. Furthermore, the final algorithm remains a ‘black box’ which can be difficult for a non-specialist to interrogate. However, all of these steps have been tested within Computer Science and other fields over many years, and it is reasonable to assume that the final set of results is reliable. The same assumptions and arguments are applicable to the use of the LDA algorithm for topic modelling.

5.6 Testing and Evaluating Topic Modelling

In the absence of a fixed number of categories at the outset, arriving at the optimal set of topics is challenging. Using the training set of 294 blogs, ten topics appears to be too few, forty-five too many. Finding the right number is computationally intensive and not always conclusive. A ‘grid search’ which calculates the log-likelihood and perplexity scores per word is one way of computing the number of topics. Ideally, one of the lines in Figure 5.5 should increase in gradient before dropping, the point at which that happens indicating the optimal number of topics; however, this did not occur in any combination of *topics* from 5 to 45. In the absence of any better model, the best method is to use a heuristic approach and increase the number of topics, visualising the result each time using the distance map, until a clearly-differentiated and reasonable number of topics is reached.

There are two main libraries used to perform LDA on a set of data: one is provided by Scikit-learn[6] (which also provides the classification algorithms referred to above); the other is Gensim[7]. While the method for calculating LDA is - or should be - the same, the parameters that are set ‘out of the box’ for each are not the same. Scikit-learn has many more parameters set as ‘default’, and so the first step is to equalise the Gensim LDA model as far as possible with this. Two parameters were set: ‘iterations = 30’ and ‘passes = 10’. These set the maximum number of iterations through the entire corpus to 30, with 10 passes through the corpus during the training phase. For both, the ‘random_state = 100’ was stipulated to ensure reproducibility. The full code is reproduced here https://github.com/sh9g14/PhDCode.

The number of topics was set at 20. With over 7,000 blogs in the data, it is likely that the number of topics discussed will be diverse. The aim is to find the largest number of topics that can be labelled coherently from the top words in each topic. It is expected that this will vary from year to year when this exercise is carried out on the final data,

[6] https://scikit-learn.org/stable/index.html
[7] https://radimrehurek.com/gensim/

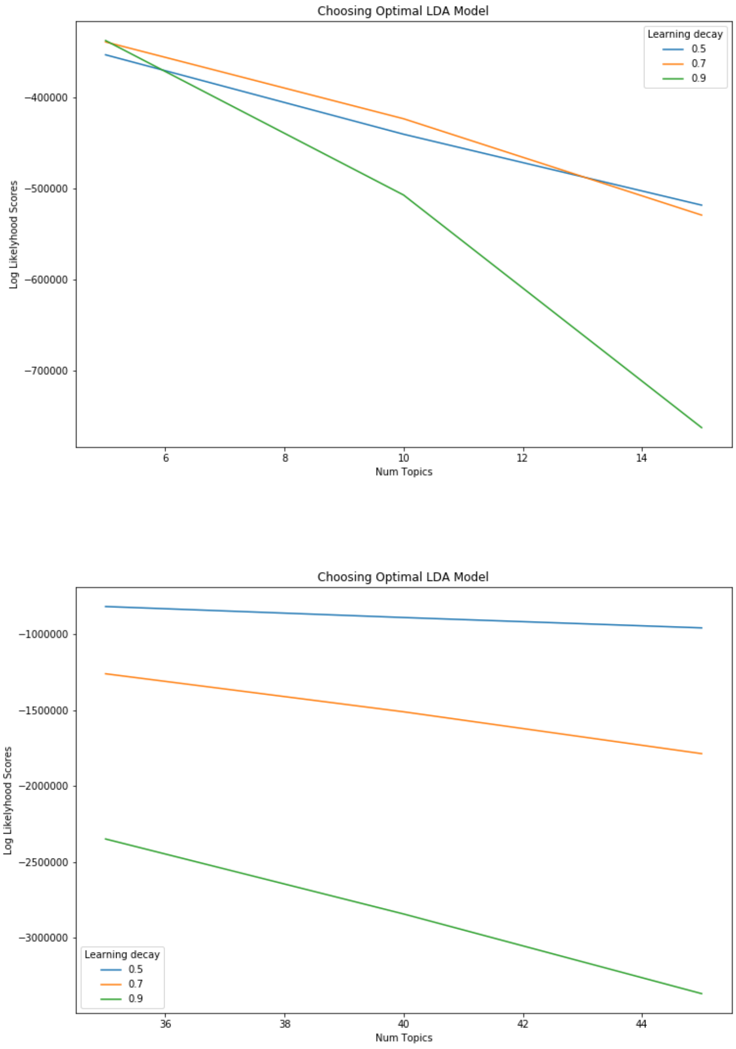
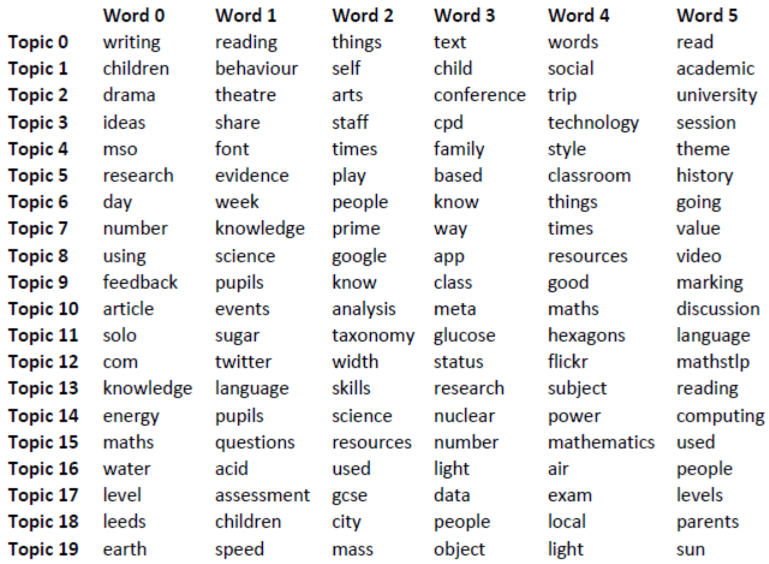


Figure 5.5: Gridsearches: *topics*=5-15; *topics*=30-45.

although it is also anticipated that there will be some coherent topics that appear in blogs each year.

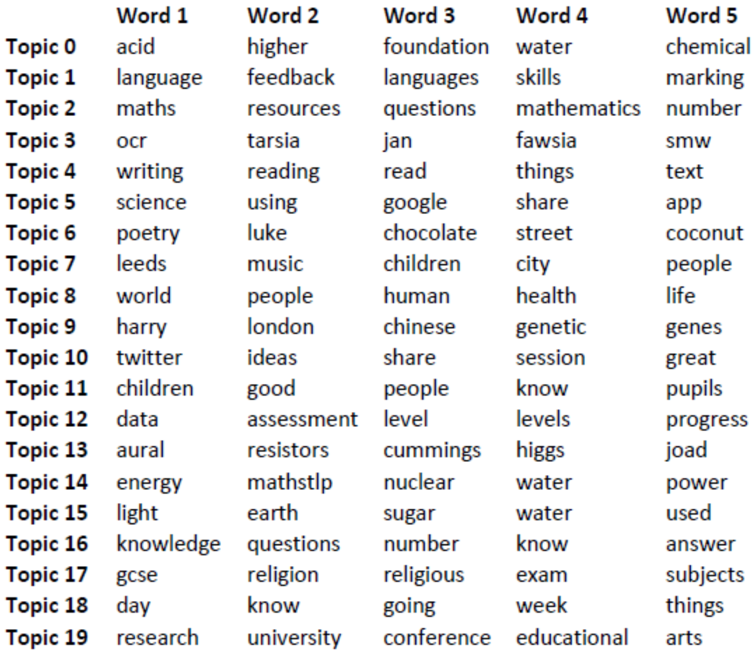
The top 5 words for each topic are shown in figure 5.6. The corresponding top 5 words using the Gensim model are shown in figure 5.7. The results from both of these providers were generated using a CV matrix.

The final step is to compare results using a word matrix calculated using TFIDF before generating the topic model. Figure 5.8 illustrates the results using Gensim. Most of the topics overlap completely, rendering this method ineffective, however the TFIDF model was used with no parameter adjustments. It is possible that with some experimentation the results could be improved.



	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	writing	reading	things	text	words	read
Topic 1	children	behaviour	self	child	social	academic
Topic 2	drama	theatre	arts	conference	trip	university
Topic 3	ideas	share	staff	cpd	technology	session
Topic 4	mso	font	times	family	style	theme
Topic 5	research	evidence	play	based	classroom	history
Topic 6	day	week	people	know	things	going
Topic 7	number	knowledge	prime	way	times	value
Topic 8	using	science	google	app	resources	video
Topic 9	feedback	pupils	know	class	good	marking
Topic 10	article	events	analysis	meta	maths	discussion
Topic 11	solo	sugar	taxonomy	glucose	hexagons	language
Topic 12	com	twitter	width	status	flickr	mathstlp
Topic 13	knowledge	language	skills	research	subject	reading
Topic 14	energy	pupils	science	nuclear	power	computing
Topic 15	maths	questions	resources	number	mathematics	used
Topic 16	water	acid	used	light	air	people
Topic 17	level	assessment	gcse	data	exam	levels
Topic 18	leeds	children	city	people	local	parents
Topic 19	earth	speed	mass	object	light	sun

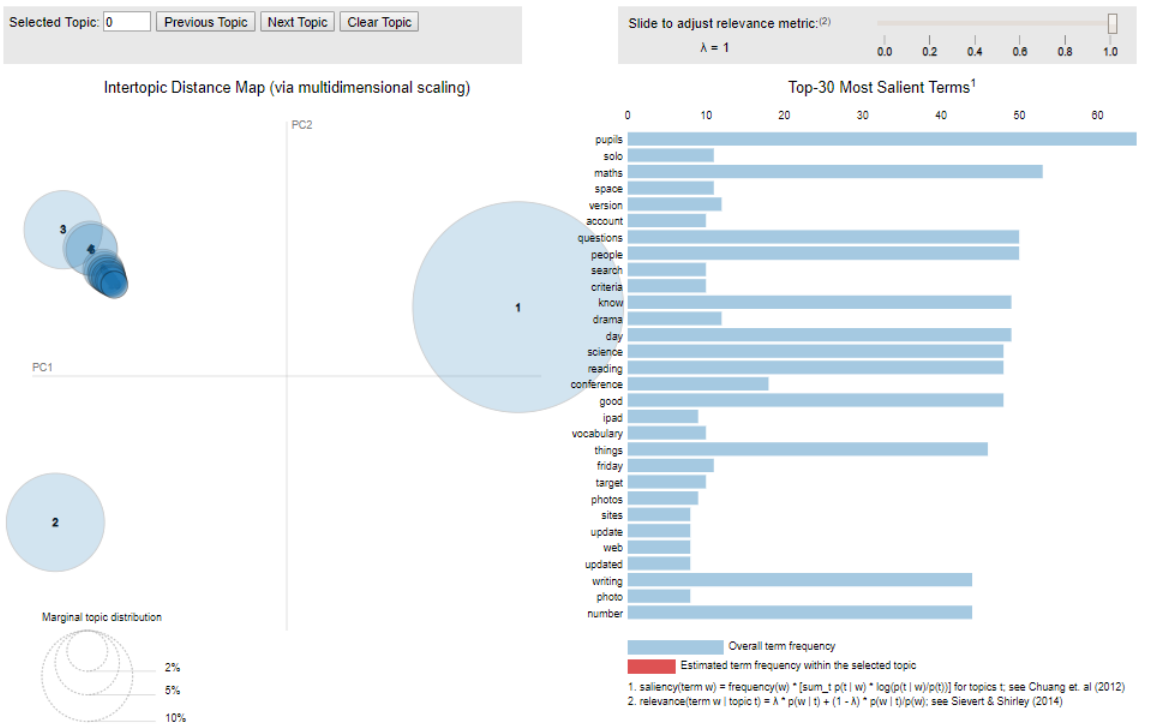
Figure 5.6: Top 5 words per topic using Scikit-learn, *topics*=20



	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	acid	higher	foundation	water	chemical
Topic 1	language	feedback	languages	skills	marking
Topic 2	maths	resources	questions	mathematics	number
Topic 3	ocr	tarsia	jan	fawsia	smw
Topic 4	writing	reading	read	things	text
Topic 5	science	using	google	share	app
Topic 6	poetry	luke	chocolate	street	coconut
Topic 7	leeds	music	children	city	people
Topic 8	world	people	human	health	life
Topic 9	harry	london	chinese	genetic	genes
Topic 10	twitter	ideas	share	session	great
Topic 11	children	good	people	know	pupils
Topic 12	data	assessment	level	levels	progress
Topic 13	aural	resistors	cummings	higgs	joad
Topic 14	energy	mathstlp	nuclear	water	power
Topic 15	light	earth	sugar	water	used
Topic 16	knowledge	questions	number	know	answer
Topic 17	gcse	religion	religious	exam	subjects
Topic 18	day	know	going	week	things
Topic 19	research	university	conference	educational	arts

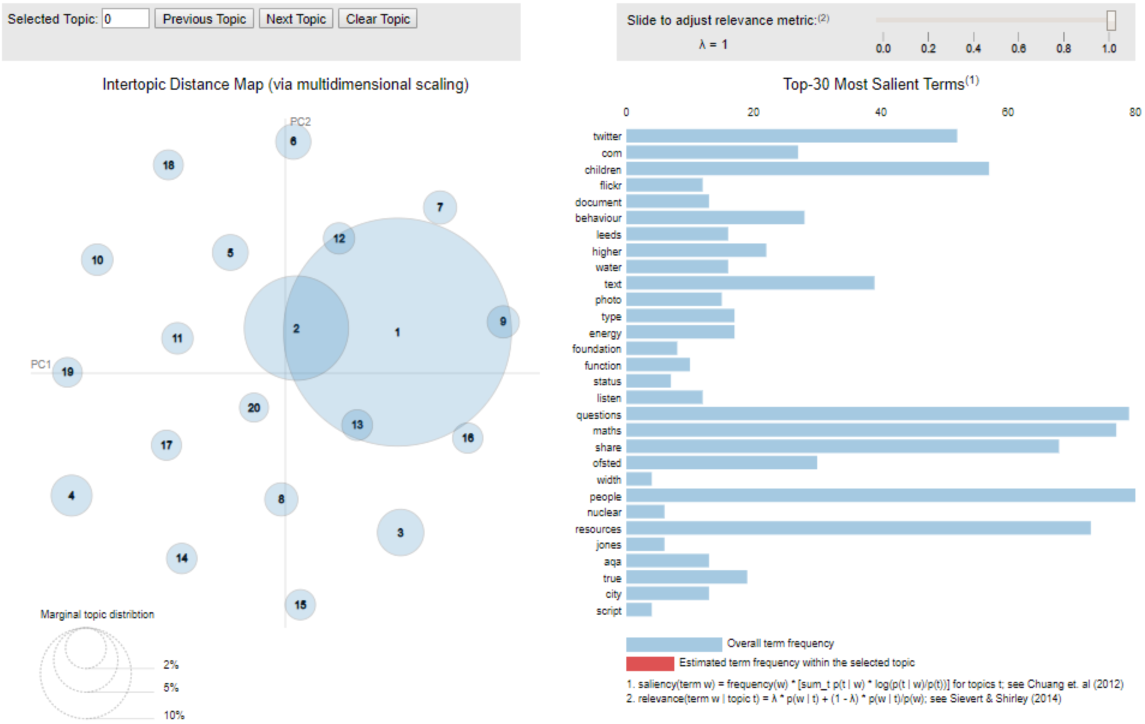
Figure 5.7: Top 5 words per topic using Gensim, *topics*=20

The topics appear to be much better separated by Scikit-learn (figure 5.9). However, if we take a closer look at the top 5 words per topic (figure 5.10), some of the top 5 words make deciding a topic label challenging e.g. topics 7 and 16. Weighting word counts according to the length of the document brings the words used least frequently to the fore, and while several papers discussed in Chapter 2 suggest that using TFIDF is an effective way of clustering documents using, for example, k-means, it produced poorer accuracy scores when testing a range of classifiers (see above), and is less satisfactory than using a simple count vector here.



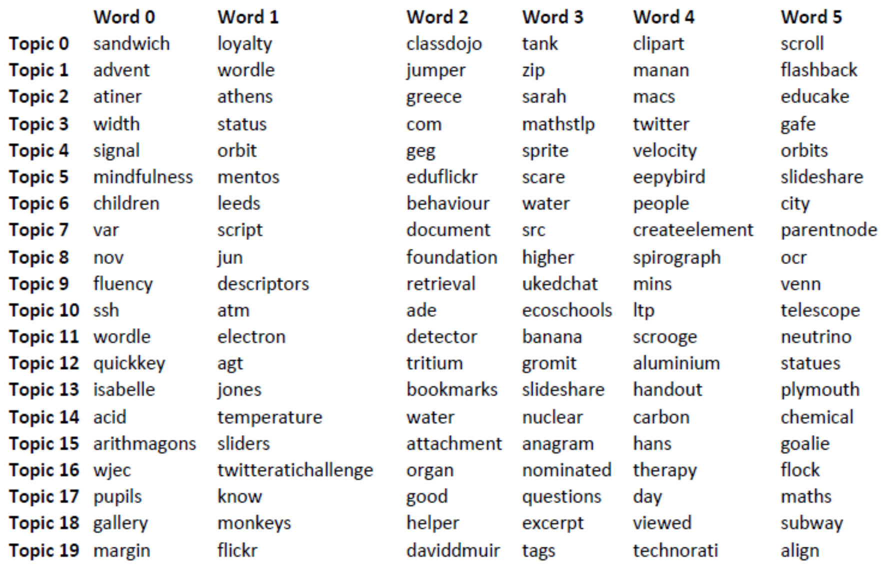
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.8: Topic Modelling using Gensim and TFIDF, *topics*=20



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.9: Topic Modelling using Scikit-learn and TFIDF, *topics*=20



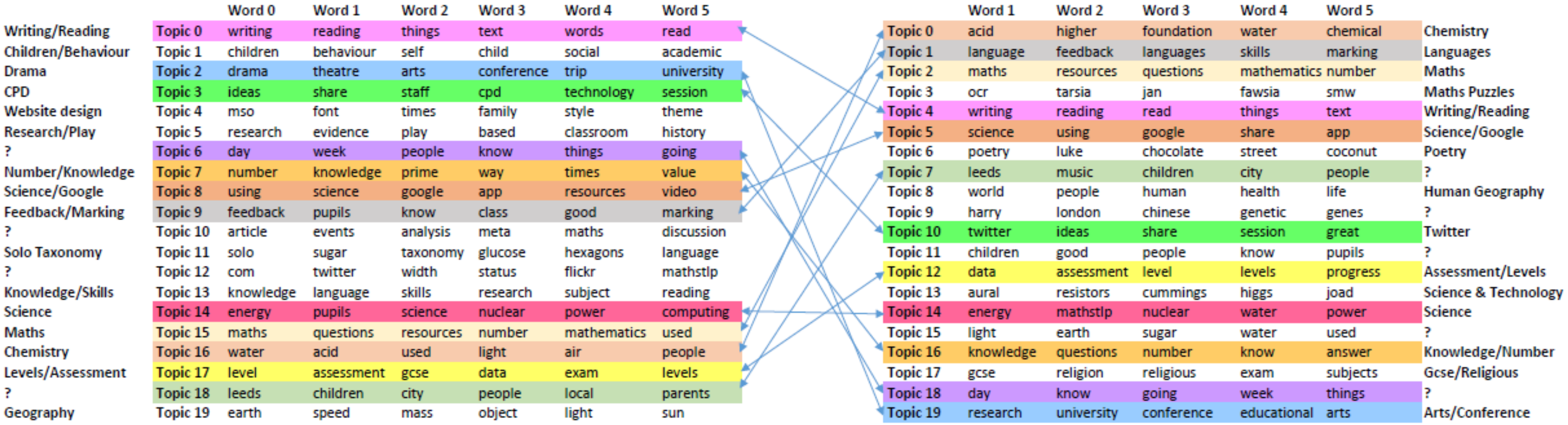
	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 0	sandwich	loyalty	classdojo	tank	clipart	scroll
Topic 1	advent	wordle	jumper	zip	manan	flashback
Topic 2	atiner	athens	greece	sarah	macs	educake
Topic 3	width	status	com	mathstlp	twitter	gafe
Topic 4	signal	orbit	geg	sprite	velocity	orbits
Topic 5	mindfulness	mentos	eduflickr	scare	eepybird	slideshare
Topic 6	children	leeds	behaviour	water	people	city
Topic 7	var	script	document	src	createelement	parentnode
Topic 8	nov	jun	foundation	higher	spirograph	ocr
Topic 9	fluency	descriptors	retrieval	ukedchat	mins	venn
Topic 10	ssh	atm	ade	ecoschools	ltp	telescope
Topic 11	wordle	electron	detector	banana	scrooge	neutrino
Topic 12	quickkey	agt	tritium	gromit	aluminium	statues
Topic 13	isabelle	jones	bookmarks	slideshare	handout	plymouth
Topic 14	acid	temperature	water	nuclear	carbon	chemical
Topic 15	arithmagons	sliders	attachment	anagram	hans	goalie
Topic 16	wjec	twitteratichallenge	organ	nominated	therapy	flock
Topic 17	pupils	know	good	questions	day	maths
Topic 18	gallery	monkeys	helper	excerpt	viewed	subway
Topic 19	margin	flickr	daviddmuir	tags	technorati	align

Figure 5.10: Top 5 words per topic using Scikit-learn and TFIDF, *topics*=20

Both sets of word counts have been labelled, colour coded, and common topics indicated by arrows in Figure 5.11. There are some differences between the two models as well as similarities. However, there is little to choose between them. Both present some categories that are difficult to label, although it is clear that some additional words could be treated at stopwords, such as ‘use’ and ‘using’. Some words could be replaced i.e. ‘mathematics’ could be replaced with ‘maths’. The Gensim model has 5 topics that are hard to label, the Scikit-learn model has 4. Both models identified two small topics - topic 4 in the Scikit-learn model identified at least one blog post discussing web site design, Gensim identified a topic relating to maths puzzles attributable to one blogger. However, neither found *both* topics. The decision as to which to use therefore comes down to practicalities: which library is less computationally expensive, i.e. runs faster and more efficiently on the hardware being used to process the data? In this case, it is Scikit-learn.

As mentioned previously, there is no definitive number that can be discerned that will produce a “correct” number of easily-labelled topics. However, the visual representation provided by LDAvis, together with a list of the top 10 terms per topic, present a way of adjusting the number so that the final number presents a range of topics that fill the two-dimensional space represented by LDAvis without overlapping, and produce a set of words that suggest a clear topic e.g. use of technology, reading, classroom behaviour etc. As well as using LDAvis, a two-dimensional representation of the topics was generated t-SNE (explained in more detail in Appendix B). This represents the topics using colour, providing a visual means of evaluating how well-separated (or not) they are. In fact, all the data sets could be divided into 12, 15 or 20 topics depending on the number of posts per year.

A set of tables showing the topic number, number of documents per topic, and the top 10 key words for each year are included in Appendix C.



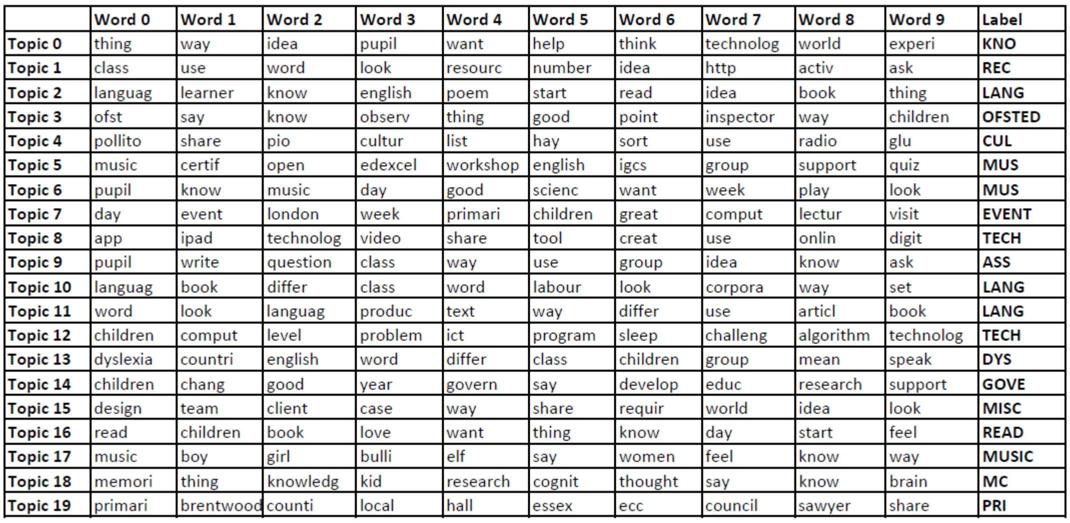
		Word 0	Word 1	Word 2	Word 3	Word 4	Word 5
Writing/Reading	Topic 0	writing	reading	things	text	words	read
Children/Behaviour	Topic 1	children	behaviour	self	child	social	academic
Drama	Topic 2	drama	theatre	arts	conference	trip	university
CPD	Topic 3	ideas	share	staff	cpd	technology	session
Website design	Topic 4	mso	font	times	family	style	theme
Research/Play	Topic 5	research	evidence	play	based	classroom	history
?	Topic 6	day	week	people	know	things	going
Number/Knowledge	Topic 7	number	knowledge	prime	way	times	value
Science/Google	Topic 8	using	science	google	app	resources	video
Feedback/Marking	Topic 9	feedback	pupils	know	class	good	marking
?	Topic 10	article	events	analysis	meta	maths	discussion
Solo Taxonomy	Topic 11	solo	sugar	taxonomy	glucose	hexagons	language
?	Topic 12	com	twitter	width	status	flickr	mathstlp
Knowledge/Skills	Topic 13	knowledge	language	skills	research	subject	reading
Science	Topic 14	energy	pupils	science	nuclear	power	computing
Maths	Topic 15	maths	questions	resources	number	mathematics	used
Chemistry	Topic 16	water	acid	used	light	air	people
Levels/Assessment	Topic 17	level	assessment	gcse	data	exam	levels
?	Topic 18	leeds	children	city	people	local	parents
Geography	Topic 19	earth	speed	mass	object	light	sun

	Word 1	Word 2	Word 3	Word 4	Word 5	
Topic 0	acid	higher	foundation	water	chemical	Chemistry
Topic 1	language	feedback	languages	skills	marking	Languages
Topic 2	maths	resources	questions	mathematics	number	Maths
Topic 3	ocr	tarsia	jan	fawsia	smw	Maths Puzzles
Topic 4	writing	reading	read	things	text	Writing/Reading
Topic 5	science	using	google	share	app	Science/Google
Topic 6	poetry	luke	chocolate	street	coconut	Poetry
Topic 7	leeds	music	children	city	people	?
Topic 8	world	people	human	health	life	Human Geography
Topic 9	harry	london	chinese	genetic	genes	?
Topic 10	twitter	ideas	share	session	great	Twitter
Topic 11	children	good	people	know	pupils	?
Topic 12	data	assessment	level	levels	progress	Assessment/Levels
Topic 13	aural	resistors	cummings	higgs	joad	Science & Technology
Topic 14	energy	mathstlp	nuclear	water	power	Science
Topic 15	light	earth	sugar	water	used	?
Topic 16	knowledge	questions	number	know	answer	Knowledge/Number
Topic 17	gcse	religion	religious	exam	subjects	Gcse/Religious
Topic 18	day	know	going	week	things	?
Topic 19	research	university	conference	educational	arts	Arts/Conference

Figure 5.11: Topics and Top 5 Words from Scikit-learn and Gensim

5.7 Topic Labelling

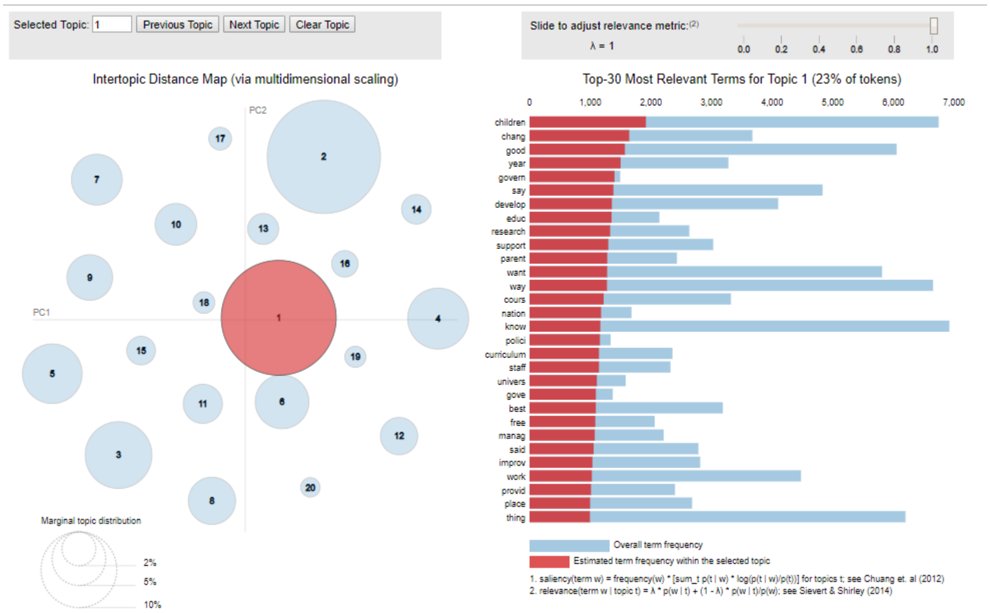
The algorithm used for topic modelling - LDA - produces a list of the top-ranked key words for each topic. Agrawal et al. (2018) have argued that the reliability of each key word accurately representing the underlying topic declines after the fifth word. Tables for each year are are included in Appendix C. The table for 2013 is shown in Figure 5.12. Topic 14 according to the table is labelled 'GOVE', yet the only indication that this might be an appropriate label is 'govern'. However, if we look at the LDAVis chart for 2013 (see Figure 5.13) for topic 1 (LDAVis ranks the topics by the number of documents in each topic i.e. the topic with the most blog posts allocated is number 1) we can see that the word 'gove' appears just over halfway down. By adjusting the slider and increasing the relevancy of the term to the topic, 'gove' moves up the list to fifth place (see Figure 5.14). 'Govern' and 'polici' are now prominent, making these words appear more relevant to the underlying topic.



	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Label
Topic 0	thing	way	idea	pupil	want	help	think	technolog	world	experi	KNO
Topic 1	class	use	word	look	resourc	number	idea	http	activ	ask	REC
Topic 2	languag	learner	know	english	poem	start	read	idea	book	thing	LANG
Topic 3	ofst	say	know	observ	thing	good	point	inspector	way	children	OFSTED
Topic 4	pollito	share	pio	cultur	list	hay	sort	use	radio	glu	CUL
Topic 5	music	certif	open	edexcel	workshop	english	igcs	group	support	quiz	MUS
Topic 6	pupil	know	music	day	good	scienc	want	week	play	look	MUS
Topic 7	day	event	london	week	primari	children	great	comput	lectur	visit	EVENT
Topic 8	app	ipad	technolog	video	share	tool	creat	use	onlin	digit	TECH
Topic 9	pupil	write	question	class	way	use	group	idea	know	ask	ASS
Topic 10	languag	book	differ	class	word	labour	look	corpora	way	set	LANG
Topic 11	word	look	languag	produc	text	way	differ	use	articl	book	LANG
Topic 12	children	comput	level	problem	ict	program	sleep	challeng	algorithm	technolog	TECH
Topic 13	dyslexia	countri	english	word	differ	class	children	group	mean	speak	DYS
Topic 14	children	chang	good	year	govern	say	develop	educ	research	support	GOVE
Topic 15	design	team	client	case	way	share	requir	world	idea	look	MISC
Topic 16	read	children	book	love	want	thing	know	day	start	feel	READ
Topic 17	music	boy	girl	bulli	elf	say	women	feel	know	way	MUSIC
Topic 18	memori	thing	knowledg	kid	research	cognit	thought	say	know	brain	MC
Topic 19	primari	brentwood	counti	local	hall	essex	ecc	council	sawyer	share	PRI

Figure 5.12: Topic Keywords and Labels: 2013

Adjusting the LDAVis chart for each topic can be a useful technique when a relevant word shows in the top 30 most relevant terms, but it does not necessarily reveal a clear topic when none is present. Topic 15 of Figure 5.12 is labelled 'MISC' (miscellaneous) because no clear education-related topic was apparent. Using this technique to discover meaningful Edu-related topic labels does not imply that the researcher altered settings until a suitable word was "revealed" - if an appropriate word was not apparent from the top 30, or was towards the end of the list, the topic was labelled 'MISC'. In short, where adjusting the relevancy of words to the topic provided a clearer representation of an underlying topic, this was used; otherwise the topic was left unlabelled. As stated in Murakami et al., "...the decision on how many topics a corpus will be deemed to contain is a subjective one and

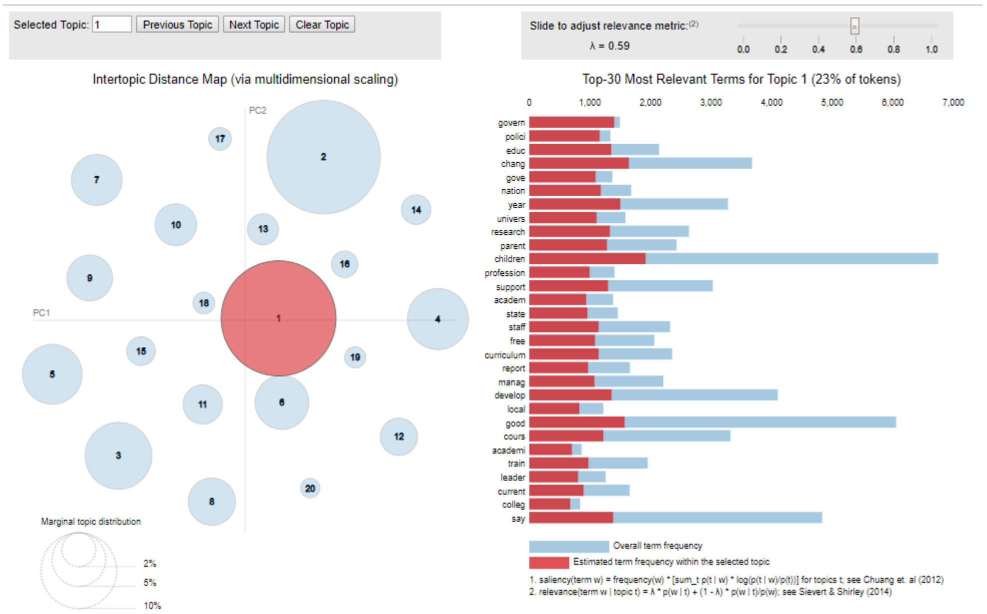


Selected Topic: 1 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.13: LDAVis Chart for 2013, topic 1 highlighted



Selected Topic: 1 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2) λ = 0.59

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.14: LDAVis Chart for 2013, topic 1 highlighted and word saliency adjusted

the answer may be defended on the grounds of usefulness but not on the grounds of accuracy"(Murakami et al., 2017, p.250).

A final way to check how well-separated the topics are is to construct a 2D image of the data. There are, of course, different algorithms available to achieve this, but one of the most widely used is t-SNE (t distribution-Stochastic neighbourhood embedding). A t-SNE graph is provided for every year of blog data, but the graph for 2013 is shown in Figure 5.15. The top 5 words for each topic are included in the graph, although the colours do not correspond with any previous figures. Generally, the clusters are fairly well separated, although a closer look at the colours will reveal that there are some blog posts that are some way from the cluster to which they have been assigned. Nevertheless, given that there are over 6,000 blog posts for the year 2013, and some could legitimately belong to two or more topics, the result is acceptable. A t-SNE graph is shown for every year in Appendix B.

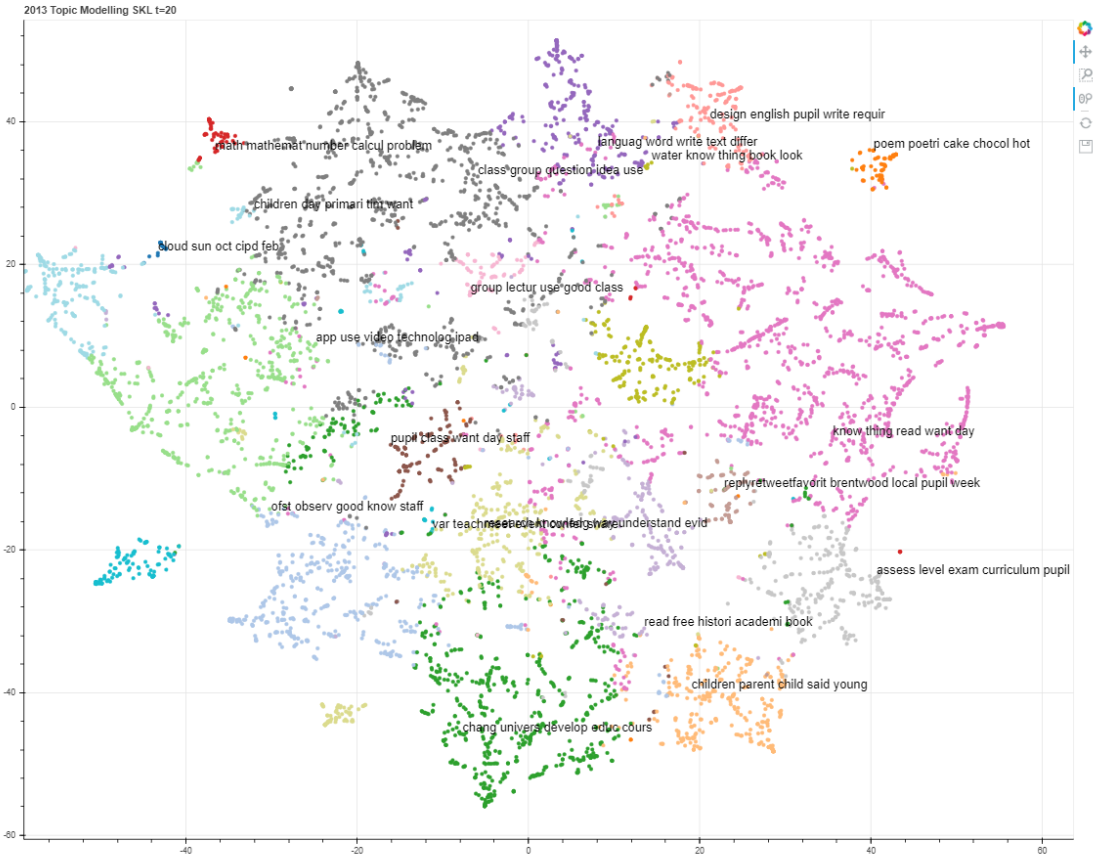


Figure 5.15: 2D Representation of 20 Topics generated by LDA, 2013

5.8 Final Notes on Classifiers and Clustering Algorithms

A classifier applies one of a finite number of labels to a set of documents, using rules 'learned' during a training phase. Topic modelling groups documents with no *a priori* assumptions. The only research paper to take this approach was Galyardt et al. (2009), although the authors took their data from one blog-service site. It is interesting to note that they did not remove any stopwords, and so their results retained verbs such as 'can', 'has' and 'will'. Consequently, a category such as 'Technical Help Seeking' has more of a thematic-feel to it than 'Foreign language'. As the site was set up for teachers seeking help for IT-related issues, most of the 20 categories are linked by IT.

The category labels used with the mNB classifier also have a thematic look. Words such as ‘positioning’ and ‘soapboxing’ are verbs, although as labels they are acting as nouns. In contrast, topic modelling produces a set of labels (or at least, the researcher has deduced the labels) which are nouns e.g. ‘poetry’ or ‘Ofsted’; however the blogs in the topics will also be thematic, and the categories will also contain blogs that relate to broader topics.

Topic modelling itself has some inherent challenges. Agrawal et al. (2018) note that there are several factors that can impact the reliability and accuracy of the results. These are ‘sampling’, learner’, ‘evaluation’ and ‘order’ bias. These have important implications for this entire research, and will be discussed in more detail below. Suffice to say for the moment that variations in the data going in to the LDA algorithm, the parameters of the algorithm itself and the way the results are evaluated all impact the results are presented in the previous Chapter. LDA is a *probalistic* algorithm that is ‘non-deterministic’ i.e. it is based on the *probability* that a document belongs to one particular topic over another, and *non-deterministic* in that it will not give exactly the same results each time it is run (although the variation can be limited if the recommendations made in Agrawal et al. (2018) are followed).

The success of any classifier or clustering algorithm also depends on the shape and distribution of the underlying data. For example, imagine each point in Figure 5.16 represents a document in a set of data. Its placement in this two-dimensional space has been calculated using cosine similarity (see Chapter 2, Section 2.5.2). The data splits easily into four groups or clusters. How would an algorithm split the clusters as per Figure 5.17? Now imagine data plotted in 3D space (see Figure 5.18[8]) and the scale of the problem becomes apparent, as it becomes a decision as to where best create a boundary between clusters of data. Both Figures 5.16 and 5.17 have around 200 data points. The data set used in this research equates to a data point for *each unique word (or token)*. There are two possible undesirable outcomes when applying a classification algorithm to unlabelled data: overfitting or underfitting. Overfitting occurs when the algorithm learns the noise and detail in the training data too well to generalise when it comes to the unlabelled data. Underfitting is the opposite problem, as the algorithm can neither model the training data, nor generalise over the unlabelled data. Plotting such a large data set to have a look at the underlying structure is infeasible, and therefore the only way to assess how best to approach applying a structure is to judge by the results of a range of algorithms. All the algorithms *except* mNB (see Chapter 3, section 3.4.3) make boundary decisions when classifying the underlying data, which would suggest that this is why it returned the most accurate score compared with the others.

As well as the shape and distribution of the data, we must also remember that both mNB and LDA convert the words in the corpus into tokens before eventually converting the tokens into word count vectors. The underlying *meaning* of the words, the sense

[8] https://matplotlib.org/gallery/mplot3d/scatter3d.html#sphx-glr-gallery-mplot3d-scatter3d-py

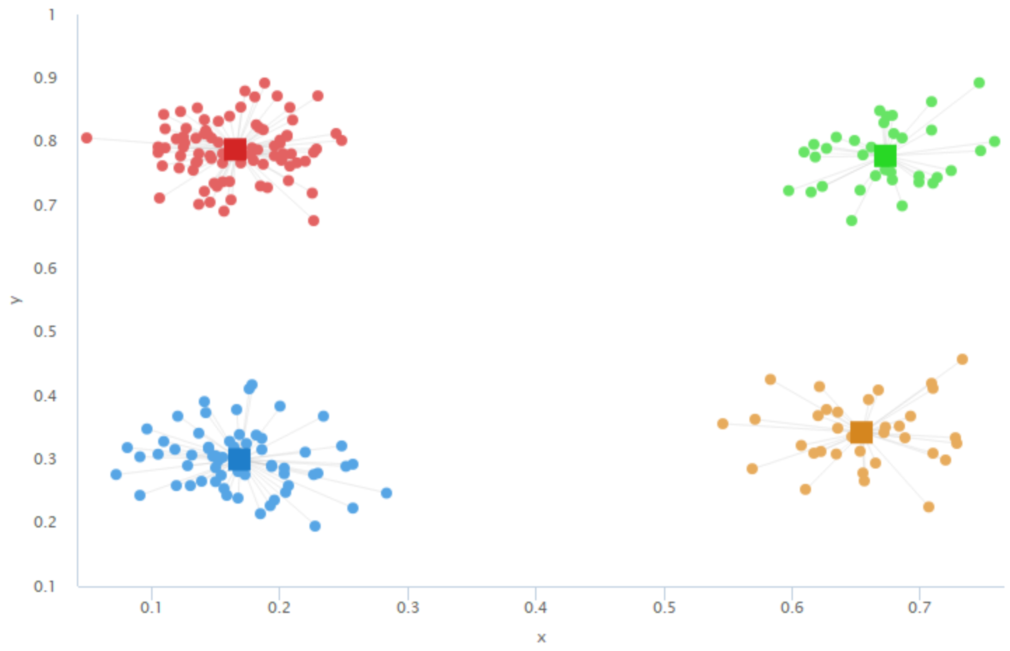


Figure 5.16: Data Distribution Example 1

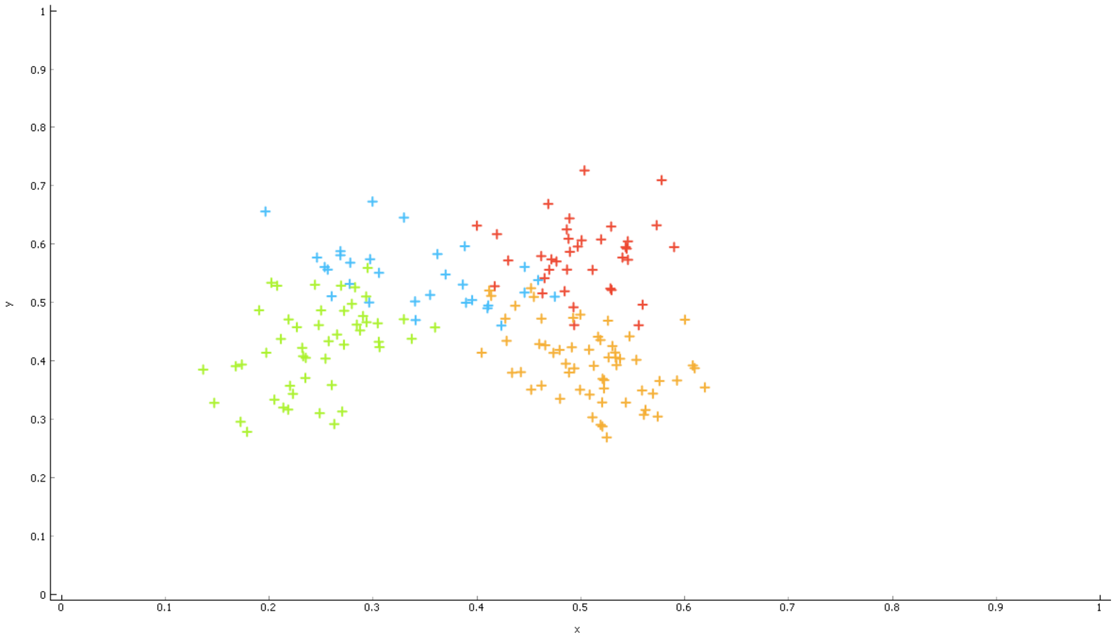


Figure 5.17: Data Distribution Example 2

created by the construction of language is entirely disregarded. The fact that mNB can do this and yet still return an accuracy score of 62% is possibly remarkable, especially when the categories are themed rather than discussing a concrete topic such as 'history' or 'Ofsted', although the results have not been evaluated by the researcher and are taken as presented. The logic, though, is clear: when looking for the *what*, certain words will occur more frequently than others and can be grouped around the (or 'a') *what*.

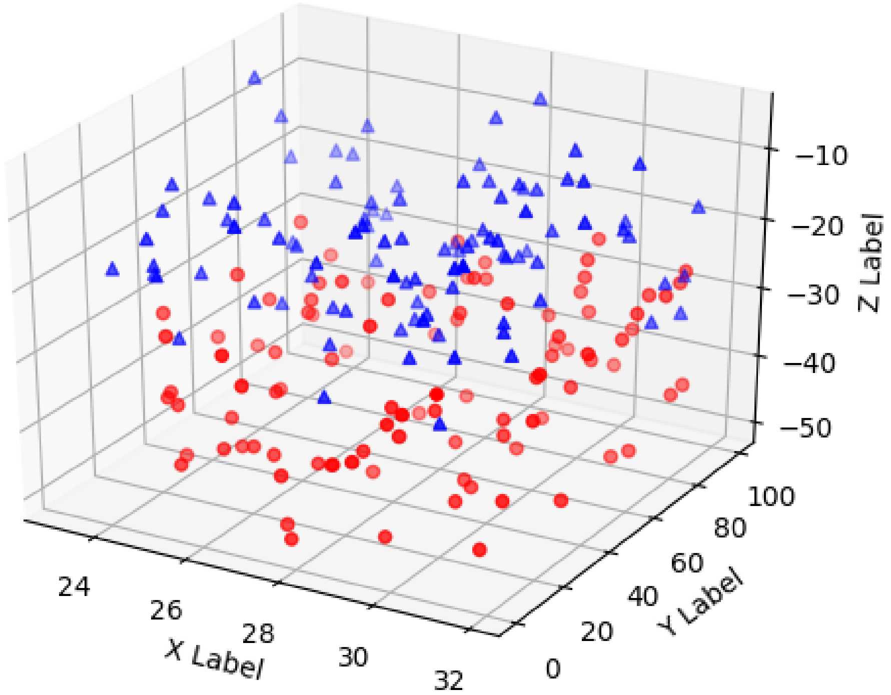


Figure 5.18: Data Shape and Distribution Example 3

5.9 The Relationship Between Categories and Topics

The Introduction section of this thesis mentioned why it was important to examine existing research into the subject of *what* the Edu-community blogs about, as well as taking an *a priori* topic modelling based approach (page 10). An examination of the results - discussed above - shows that there is almost no obvious link between the categories that emerged from the existing literature, and the topics produced by topic modelling.

The following Chapter will present the conclusions that can be drawn from this research, based on the results of the classifier and topic modelling, together with limitations and future work.

Chapter 6

Conclusions and Future Work

6.1 Introduction

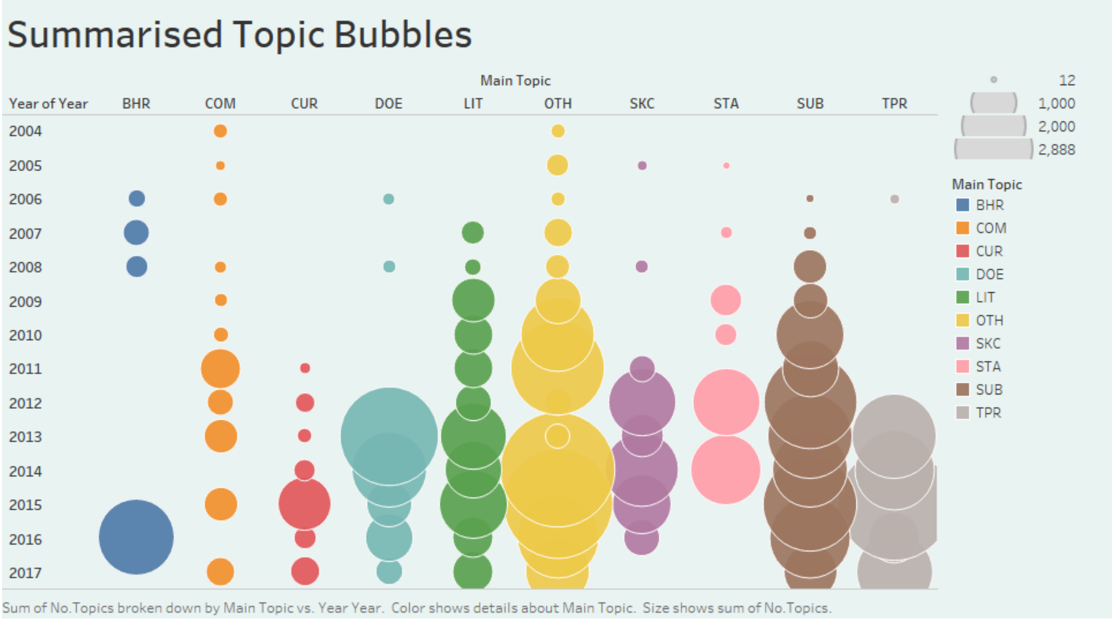
In this final Chapter, we return to the Research Questions and address them in turn. A considerable amount of thought (and work) has gone into the harvesting and preparation of the blog posts, which represent as far as possible the voices of the Edu-community. During the course of this research many more voices have become apparent - but nevertheless it is a collection of the blogs of many hundreds of Edu-bloggers, some of whom will have been writing to express opinions they might hope will be heard, others will be writing to help colleagues out with ideas and resources. Some will do both, or neither of these things; they may have some other aim.

6.2 Research Question 1: What kind of subject matter, themes or issues have been discussed by the Edu-community since the arrival of the first available Edu-blog?

Figures 4.1 and 6.4 both illustrate the the 'subject matter, themes or issues discussed by the Edu-community'. However, given that the classifier (figures 4.1 and 4.2) does not tell us anything new, the most relevant results are revealed by topic modelling. Figure 6.4 shows the full range of the topics that have occupied the Edu-Community since 2004. There are 53 discrete topics in total, showing that the community has discussed a wide range of themes, issues and subjects. Some are specific such as 'Maths', others are broader such as 'Teaching Practice'. It is very difficult to discuss *what* the community has discussed without also discussing changes over time, and therefore both questions will be addressed together.

6.3 Research Question 2: Has the subject matter, themes or issues have been discussed by the Edu-community changed since the arrival of the first available Edu-blog?

The two approaches to categorising a corpus - classification and topic modelling - produce some comparable results. The topic 'resources' appears in both, and we can suggest that 'reflective practice' from the classifier model is similar to 'teaching practice' (TP). The 'other' and 'miscellaneous' categories are very different in terms of the overall size of each category. Only a small proportion of the posts were categorised as 'Other' by the classifier, compared with topic modelling. As mentioned above, the classifier will label blogs as 'other' based on what the mNB algorithm has learned from the labelled data, and while the researcher may have had in mind 'everything not about education', the algorithm cannot interpret the data in this way. In contrast, if we refer to Figure 6.1 (summarised topics) and compare it with Figure 6.2 (the results of the classifier represented using a bubble chart for ease of comparison) we can see that this topic represented a significant proportion of the blogs for every year except 2013. The significant difference here is that topic modelling provides an opportunity to label even blogs that are not about education, leaving only the unidentifiable blogs to be labelled as 'miscellaneous'.



Sum of No.Topics broken down by Main Topic vs. Year Year. Color shows details about Main Topic. Size shows sum of No.Topics.

Figure 6.1: Main Topic Bubbles Including 'Other'

Different aspects of teaching have pre-occupied the community at different times. Looking at the main topics first (see Figure 6.1), we can observe that topics discussing Educational Policy (including topics such as Ofsted, grammar schools, academies etc.

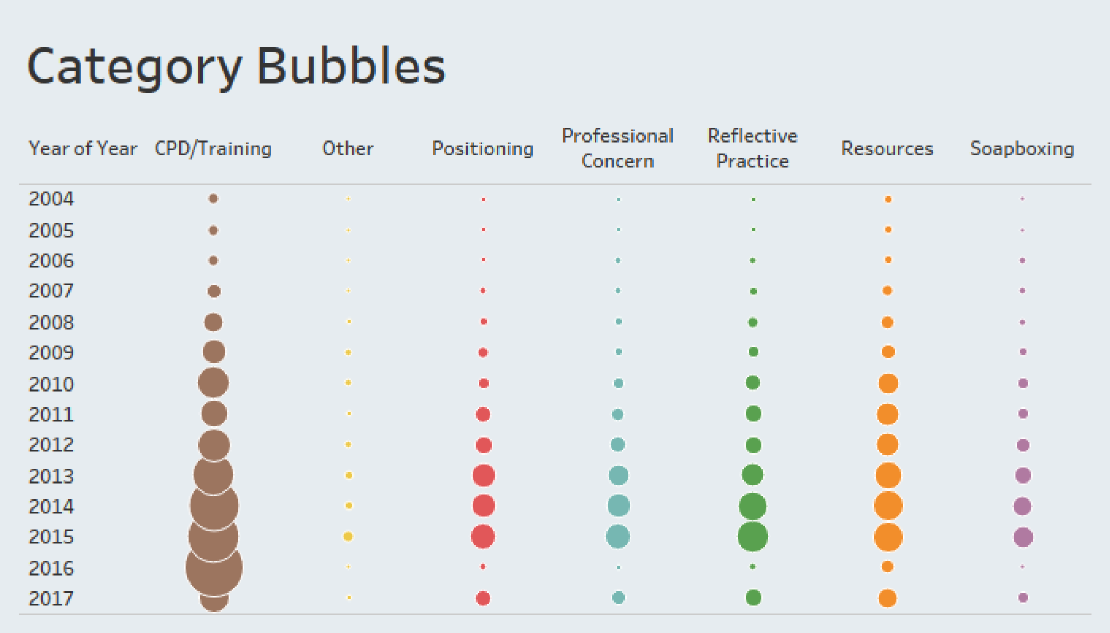


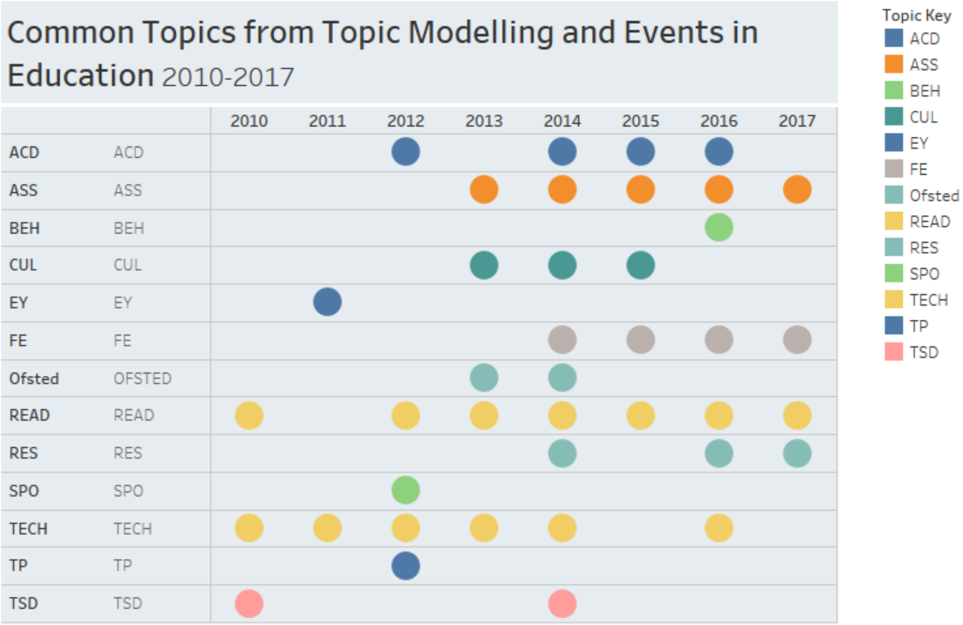
Figure 6.2: Categories represented as a bubble chart

labelled ‘DOE’) spiked in 2013, and have continued to occupy the community. Of interest to the community at the same time were matters to do with skills, knowledge and cognition (SKC), and teaching practice (TP). In 2014, blogs concerning aspects of the curriculum (which includes blogs writing about ‘Further Education’ (FE) and and ‘Higher Education’ (HE)) made a significant entry, with aspects of behaviour occupying the community in the following year.

Turning to the finer-grained topics, we can observe both an increase in the number of blogs as well as an increase in the variety of topics from around 2010. There are many reasons for this, some of which will be addressed in the discussion around the third research question, but it is likely that there are also good practical reasons such as ‘always on’ broadband available in more homes, the ease with which a blog can be set up, and a growing community wanting to engage with each other in a longer form. The main points about the blogs from this community arise when we consider the response to Research Question 3. However, it is clear from Figure 6.4 that the topics discussed *have* changed over time.

6.4 Research Question 3: Is there any link between the subjects, themes or issues discussed by the Edu-community and changes in Education policy?

Triangulating or mapping changes in Education policy (and often reports and opinion pieces written in *response* to changes) is challenging, not least because the results of topic modelling have revealed 53 different and discrete topics. A good starting point would be to look at the topics that the results of topic modelling and the 'events' as mapped in Figure 6.5 have in common. This is shown in Figure 6.3. The column on the far left shows the topic codes (slightly modified to match with the codes used for topic modelling (see table 4.2)); note that 'FE' and 'HE' are separate labels from the 'events' table, whereas only 'FE' is included in topic modelling.


Common Topics from Topic Modelling and Events in Education 2010-2017

		2010	2011	2012	2013	2014	2015	2016	2017
ACD	ACD			●		●	●	●	
ASS	ASS				●	●	●	●	●
BEH	BEH							●	
CUL	CUL				●	●	●		
EY	EY		●						
FE	FE					●	●	●	●
Ofsted	OFSTED				●	●			
READ	READ	●		●	●	●	●	●	●
RES	RES					●		●	●
SPO	SPO			●					
TECH	TECH	●	●	●	●	●		●	
TP	TP			●					
TSD	TSD	●				●			

Figure 6.3: Common Topics 2010-2017

If we compare Figure 6.3 with Figure 6.4, and if we assume that the more blogs written about a topic, the more stimulated the Edu-Community is by it, we can see evidence of impact. The first example is 'Assessment and Feedback'. This is very likely to be as a result of the various changes to the curriculum introduced by Michael Gove, as well as the impact of Ofsted inspections ('Gove', 'Ofsted' 'Teaching Practice' (TP) and 'Training and Staff Development' (TSD) also feature heavily in 2012, 2013 and 2014).

It is very difficult to triangulate the events as commented on by the web site authored by Derek Gillard, the events presented by the governments 'news and communications' web pages, and the results of topic modelling because things like an Education Act cover

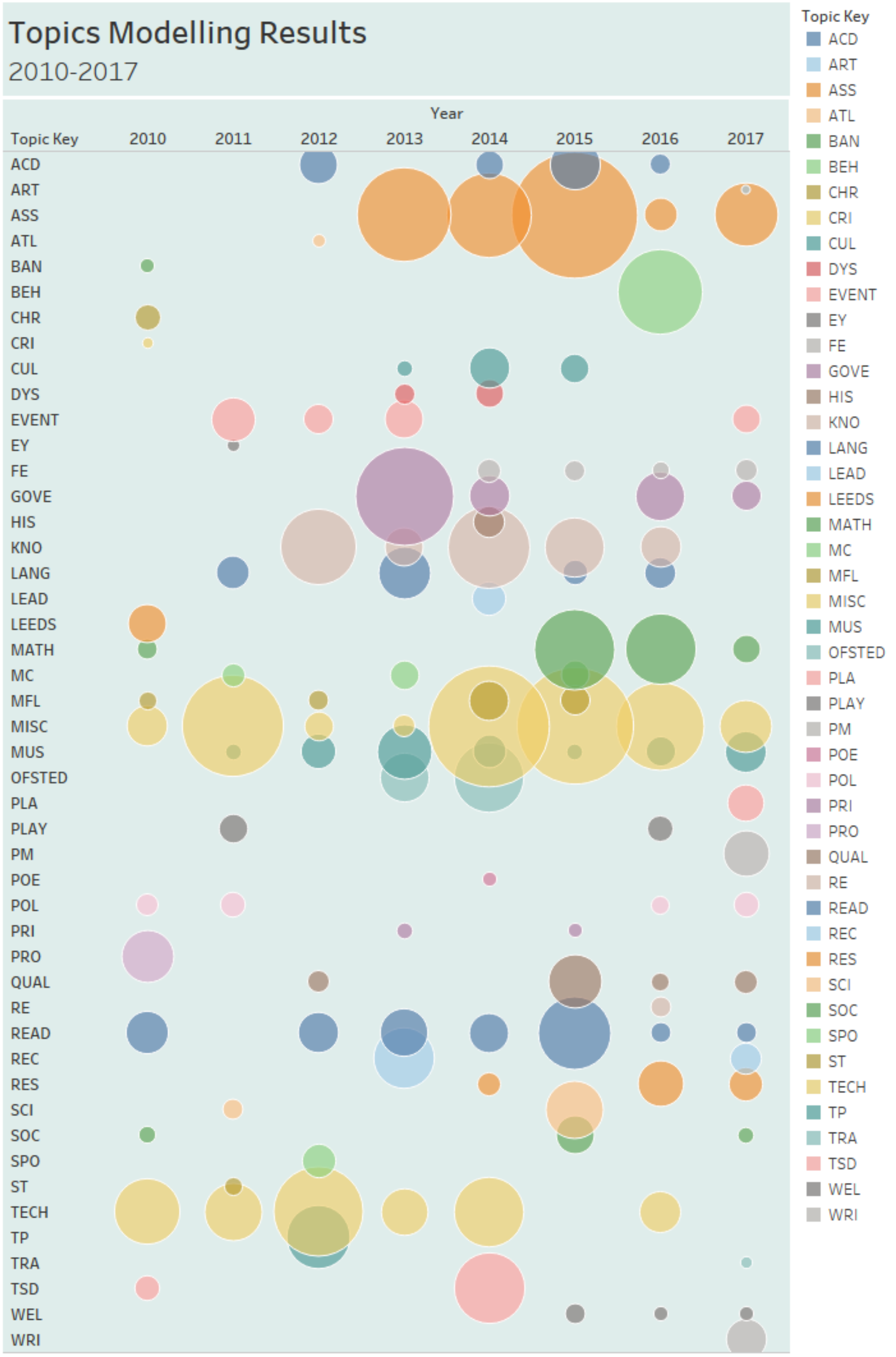


Figure 6.4: Topic Modelling Results 2010-2017

a range of issues that impact several areas of interest to the Edu-Community. Also, the impact of a particular change of policy may not be felt immediately (all governments have a habit of announcing changes at the *end* of a school year, and which may not be picked up by the Edu-blogosphere for a few weeks or months), and the language used is not always common across all sources. Nevertheless, there are clear examples of the response from the Edu-community to the changes that occured from 2010, as well as the acknowledgement of the Edu-blogosphere in chapter 5 of Peal (2015). If we begin triangulating events in Education with topics dating from 2010, we start to see some correlation.

Referring to Figure 4.5 and Chapter 19 of the 'Events' website[1], in 2010 the Academies Act introduced under the new Conservative government provided the conditions for subsequent rise in the number of academies. This is also the year that Michael Gove was appointed Secretary of State for Education. By the end of 2011, there had been a Commons Education Select Committee report into behaviour and discipline in schools; the Tickell report on Early Years Education; the Wolf Report Review on vocational Education which advocated 'academic excellence for a few students, vocational training for the rest'; the Bew report on Key Stage 2 testing; teacher training; a report focusing on literacy; a report on the government's proposals to introduce the English Baccalaureate; the Education Act which clarified a range of things including increasing schools' power to exclude students; and finally a report on a framework for a new National Curriculum.

'Behaviour' as a discrete topic does not make an appearance in the topic modelling results until 2016, but 'Teaching Practice' (TP) is an important topic in 2012 and it may well be that this topic contains discussions about the impact of behaviour and discipline. The Bew report may also have prompted discussion. In the same year, one of the topics to appear from topic modelling is 'early years', which was part of the focus for the Tickell report. However, the aspect of Gove's minister-ship that appears prominently in Figures 6.5 and 6.6 (and which the researcher remembers being discussed extensively on Twitter for most of Gove's time at the Department for Education) were changes in the National Curriculum. This is not identified as 'Curriculum' in Figure 6.4 but is represented in Figure 6.1 where it combines the results of 'Early Years' (EY) 'Further Education' (FE) and 'Primary' (PRI). It will probably be represented in almost every other topic in Figure 6.4 as 'the Curriculum' impacts all aspects of Education.

Changes to the curriculum were discussed by Dr Tina Isaacs in (Peal, 2015, Chapter 3, p.35) and by Daisy Christodoulou (in terms exams and assessment) in chapter 4, p.45 of the same publication. Things began to change. Coursework was no longer part of many exams, the traditional grading of exams using the alphabet was replaced by numbers, the 'national curriculum' as a prescriptive entity was abolished along with assessment at key stages 3 and 4 using 'levels'. Gove wanted to replace GCSEs with the EBacc (the English Baccalaureate) but this was rejected. The evidence for the effect this had on the Edu-community is almost certainly present in every blog post from the school year starting in September 2012 onwards, but is particularly noticeable in the sharp rise in the number of blogs discussing the curriculum.

[1] http://www.educationengland.org.uk/history/chapter19.html

Other evidence of blogs reflecting wider events in Education includes blogs on technology. These made a significant appearance in 2010, peaking in 2012 but remaining an important topic until the end of 2017. Technology was specifically addressed by Gove in 2012, 2013 and 2014, and was the subject of a committee report by the House of Lords in 2015 as part of 'the UKs Digital Future'.

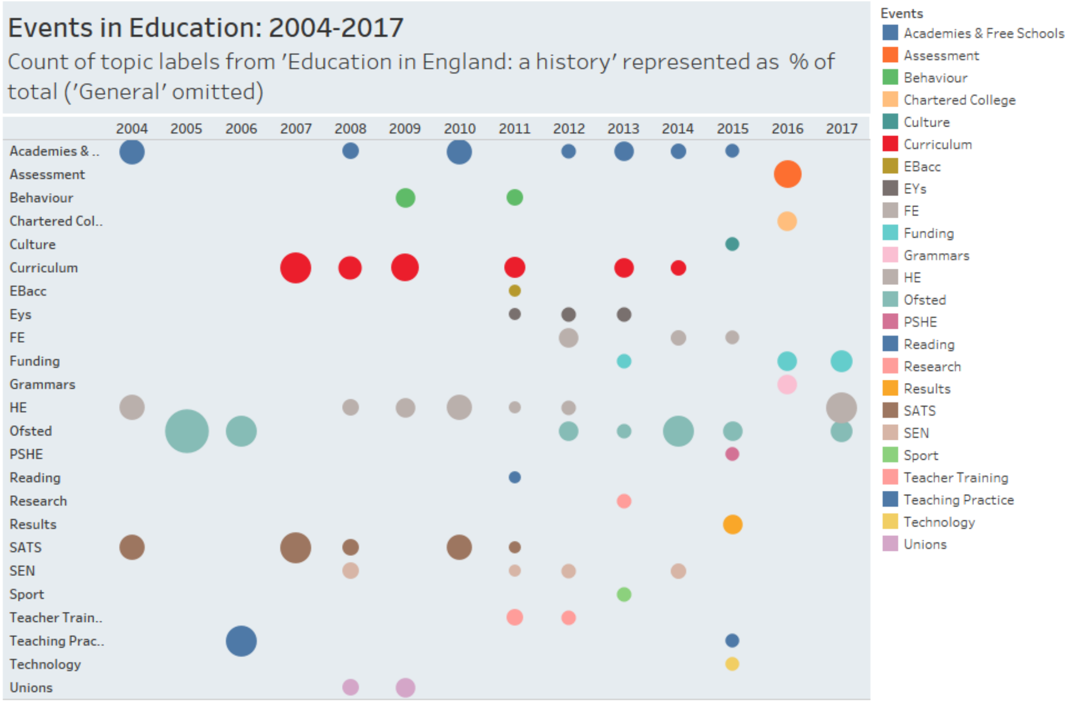


Figure 6.5: Representation of Main Events from 'Education in England'

The other topic which became a significant factor in Gove's minister-ship was Ofsted. In 2012, Ofsted introduced a new grading system for inspections. Schools rated 'good' or outstanding count convert to Academies (therefore leaving local education authority control); schools rated less than 'inadequate' were effectively viewed as failing and were encouraged to either convert to Academies, or allow themselves to be taken over by an Academy chain; schools rated 'inadequate' had to be taken over by an Adacemy sponsor. Such was the governments enthusiasm for Academies, Ofsted were accused of deliberately failing 'marginal' schools to promote Academisation. Whether this is true of not is not important here; what is certainly the case is that the Department of Education was using 'brokers' to persuade schools to become Academies in order to meet government quotas[2], as reported by 'Schools Week' in 2015. As discussed in Chapter 1, Ofsted came in for some heavy criticism for prescribing 'teaching styles', and eventually were instructed to change their inspection regime. At the beginning of 2014, Ofsted were accused of changing reports. The number of blogs mentioning Ofsted peaked in this year.

[2] https://schoolsweek.co.uk/dfe-ditches-controversial-academy-brokers-for-education-advisors/

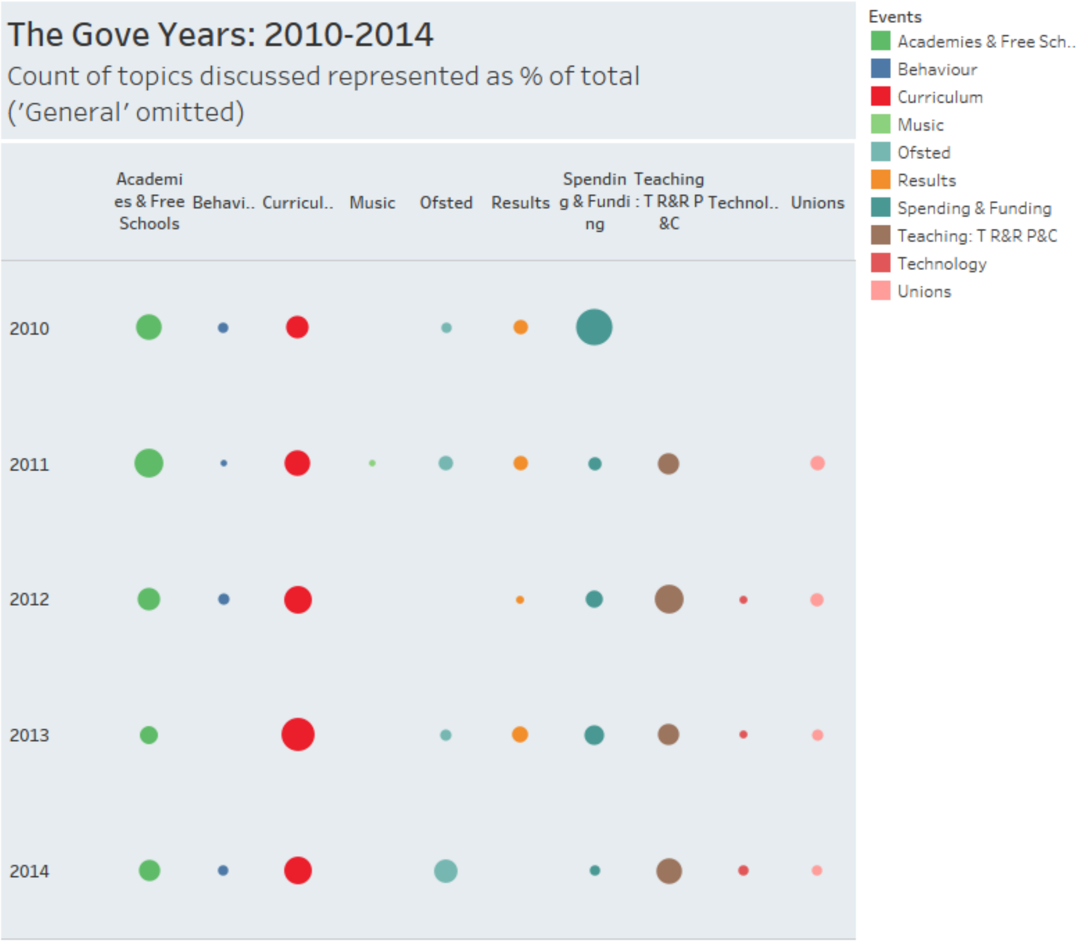


Figure 6.6: Representation of News and Communication regarding Michael Gove, published by gov.uk

The preoccupation with Academies continued in 2013 with a report criticising the overspend and an announcement from the Academies Commission (a now defunct 'think tank' set up by the RSA and the Pearson Think Tank) that Academies were 'unleashing greatness'[3]. Gove's plans to abolish GCSEs was criticised by the Commons Education Select Committee; Ofsted published new inspection arrangements; and researchED was launched by Tom Bennett. 'Research' as a topic appears in blogs from 2014.

Ofsted continued to dominate through 2015, forcing Michael Wilshaw, the then Chief Inspector of Schools in England (and the head of Ofsted) to sack some inspectors and write to all schools stating that in future lessons would no longer be graded. A new Secretary of State for Education was appointed in July. The Commons Education Select Committee said 'academies and free schools had had little or no effect on improving standards'.

[3] https://www.thersa.org/discover/publications-and-articles/reports/unleashing-greatness-getting-the-best-from-an-academised-system

In 2016, the formation of a Chartered College of Teaching was announced; ‘Attainment 8’ and ‘Progress 8’ were implemented (replacing levels); Teresa May announced that she would encourage the creation of new Grammar schools; and the National Audit Office warned of the impact of cuts to school funding. In the blogosphere, Gove was still trending but blogs about behaviour had re-surfaced. It is possible that these were connected with the general discussions still on-going about Academy and Free schools and their various admissions policies that some claimed were resulting in students being refused entry. The pressures of SATs, GCSE and A-Level results (plus Ofsted inspections still taking their toll) may also have moved the conversation towards the repercussions of poor behaviour.

At the beginning of 2017, Amanda Spielman was appointed Her Majesty’s Chief Inspector of Education, Children’s Services and Skills. Justin Greening replaced Nicky Morgan as Secretary of State for Education. No one blog topic dominated in 2017, although Gove’s name still continued to appear.

The events which had the greatest impact on the subjects of Edu-community blogs seem to have occured from 2012 onwards i.e. once the reforms introduced by Michael Gove began to bite. This is evident from Figures 4.4 and 4.3, where the bubbles representing the number of blogs in a particular topic or area increase in size significantly, reflecting proportional increase in blogs. However, if we look at Figures 4.2 and 4.1, there is no evidence of impact, which is probably due to the ‘thematic’ nature of the categories.

To conclude, then, there is evidence that the topics discussed by the Edu-blogosphere, at least from 2010 onwards, *were* in response to changes in Education policy.

6.5 Discussion

The Edu-communities’ engagement with issues and topics arising from their profession are wide and varied. This community has much to say, and many voices with which to say it. The driving force behind this research was to demonstrate that the Edu-community is not only concerned with the factors that impact on their own practice in the classroom - matters such as planning, marking and assessment for example - but are also concerned about and attempt to engage with the matters that impact the life chances of the students they teach. While many blogs have been written about Gove, Ofsted and Government policy, the results of classifying blogs according to the abstract themes revealed by existing research suggest that most of this has been done in a calm and measured way. More blogs were classified as ‘positioning’ or expressing ‘professional concern’ than ‘soapboxing’, although it should also be pointed out that the results of the classification process cannot be triangulated with the results of topic modelling.

It should be remembered that the collection of blogs used here is not an exhaustive collection of all blogs written by teachers. Some blogs are password-protected, others may never have been on the list compiled by Andrew Old to begin with. Others may not have been collected by the code used to harvest the data and so have also been missed. Also, however many bloggers *are* represented here, they are only a very small proportion of the 498,000 teachers in state-funded schools as of 2017[4]. Nor does the data represent the same bloggers year on year - bloggers come, and bloggers go, and the number of posts they write will vary. However, the data does represent the largest collection of blog posts from Edu-professionals collected so far, and reflects a recognisable picture of the community since 2004.

6.5.1 What was already known about blogging in the Edu-Community?

Part 1 of the Literature review, Chapter 2 told us what the Edu-community blogged about, although it is probably more accurate to say that is told us *how the community was doing it* rather than the content of their blogs. What we did *not* know was the details of the content - the nouns or labels we could attach to each post to *describe* the contents. We also had no sense of how those topics ebbed and flowed through the community as the years passed. Now we have a clearer sense of this, along with a clear indication that the community had much to say about the policies and government initiatives visited upon them.

6.5.2 What more does this research reveal about the Edu-blogosphere?

As well as the points made above, the classifier, using categories drawn from the literature, revealed that the most written-about topic was professional development (CPD) (see Figure 4.1). Teachers have consistently been concerned with their own growth and improvements as professional Educators. If we combine this with the number of blogs on reflective practice, these amount to around half of the total. The Edu-community represents itself as one which is wholly focused on the quality of the learning opportunities delivered to its students. This has not changed even when we consider some of the big events in Education, such as the appointment of Michael Gove as Secretary of State for Education in 2010. This is important when we consider that teachers have been criticised by Ofsted in the past for poor results. The number of blogs focused on professional matters suggests that the problem of poor results and failing schools lies elsewhere.

The application of topic modelling revealed a much broader range of categories. Fifty-three topics covered a wide-range of areas of concern to the Edu-Community that varied

[4] https://www.ethnicity-facts-figures.service.gov.uk/workforce-and-business/workforce-diversity/school-teacher-workforce/latest

in prominence year-on-year, and can be clearly linked with changes of government and the appointment of new Secretaries of State for Education. While the broader themes addressed by teachers carries on with little indication of wider goings-on - even 'soap-boxing' doesn't increase significantly during the Gove years - topic modelling reveals clear spikes in the number of blog posts on, for example, Ofsted and Gove.

The number of blogs concerning professional matters such as CPD and 'Reflective Practice' (RP) is also reflected in the results from topic modelling, although the labels are necessarily different. Figure 6.1 shows that the 'Teaching Practice' (TPR) category - which includes assessment, planning and marking, and teaching practice topics - formed a significant proportion of the Edu-blogosphere from 2013 onwards. In short, using topic modelling confirms that the subject matter, themes or issues discussed by the Edu-Community has changed over time, and is frequently in response to changes in Education policy.

6.6 Why this is important: Teacher Voice

As this research concludes, the current Conservative government have been preoccupied entirely with leading the UK out of Europe. A recent report from the National Audit Office into converting maintained schools into academies reported that the challenges faced by the government in continuing the programme were 'likely to increase'[5]. It also criticised the government for not being more "...rigorous in its scrutiny of applicants' financial sustainability and governance". Spending per pupil has fallen by 8% in real terms since 2009/10[6]. There are shortages in teacher recruitment in most subjects (Foster, 2018, p.9), and teacher retention figures continue to be poor (Foster, 2018). The most recent criticism to be levelled at schools is one of students exclusions being 'too high' and 'too easy'(Lough, 2019). In 2018, this was linked to school performance[7]; in 2019 the Mayor of London, Sadik Khan, blamed knife crime on the rise in exclusions[8]. It would appear from these headlines that teachers and other Edu-professionals are still in the firing line when things go wrong, are are still not being consulted to find solutions.

One area where the picture appears to be different is Ofsted. Sean Harford, an HMI and Ofsted's National Director, Education, was active on Twitter until very recently, and regularly engaged in conversation with the Edu-Community. This has led to Ofsted considering issues such as teacher workload, resulting in Her Majesty's Chief Inspector, Amanda Spielman, commissioning research in 2018[9]. The results of this have just been

[5] https://www.nao.org.uk/report/converting-maintained-schools-to-academies/#

[6] https://fullfact.org/education/school-spending-figures-misleading/

[7] https://www.theguardian.com/education/2018/sep/04/the-rise-in-school-exclusions-is-a-result-of-the-education-market

[8] https://www.theguardian.com/uk-news/2019/mar/07/pm-urged-to-fix-school-exclusion-system-to-tackle-knife-sadiq-khan

[9] https://educationinspection.blog.gov.uk/2018/11/30/teacher-well-being-and-workload-survey-interim-findings/

released. Ofsted also launched an Education Inspection Framework as a draft for consultation[10] and a number of individuals were recruited to provide specific feedback. This list was made public on Twitter with the permission of Harford, and included a number of practising teachers[11].

While the list of people working with Ofsted has attracted some criticism, it is nevertheless encouraging to see evidence of a government body engaging so directly with the Edu-Community.

6.7 Why this is important: Wider Benefits

The evidence of 'teacher voice' should be of much wider interest and benefit to the Edu-community itself, and other interested bodies such as journalists and policy makers. There is already evidence that the opinions of some of the more high-profile bloggers had a hand in influencing policy with regard to Ofsted (Peal, 2015); both sets of results obtained from the different approaches to the analysis should offer much valuable insight. The classification of the results produced blogs grouped according to broad themes. The theme of 'Professional Concern' would suggest much to interest the Department for Education, as this group of blogs is not what might be thought of as 'letting off steam', characterised in the blogs as 'soapboxing', but the measured and considered concern of Edu-professionals. This may be similar to the blogs grouped using topic modelling that are labelled 'Assessment and Feedback' or 'Planning and Marking' (the summarised group 'Teaching Practice', see Chapter 4, section 4.4).

In contrast, the groups suggested by topic modelling should be of interest to the Edu-practitioner. There are obvious groups such as 'Resources' and 'Behaviour' where teachers could seek advice and inspiration; however topics such as 'Teaching Practice' and 'Planning and Marking' suggest discussions that may also provide practical help for the teacher. Where an Education Minister of *TES*[12] journalist may be seeking an insight into the impact of Educational policy, teachers are more pragmatic and seek out specific topics to help them in their work.

6.8 Future Work

One of the drawbacks of any clustering algorithm is that it does ignore word meaning. Only word counts matter, as these are converted into numbers and from there the algorithm has something it can use to calculate with. This has proved surprisingly

[10] https://www.gov.uk/government/publications/education-inspection-framework-draft-for-consultation
[11] https://twitter.com/debrakidd/status/1100096829134462984
[12] Times Educational Supplement

effective, and given that humans could not agree on classifying thousands of documents into over 50 categories, I think they do a very acceptable job. What would be a useful extension would be to experiment with which particular words are the most useful when it comes to clustering a set of documents from this domain. A thorough evaluation of the results, especially the blogs labelled as 'miscellaneous' may well provide additional insights that could be used to develop more accurate labels. Using n-grams (pairs or groups of words that occur together) may have helped with topic labelling, as they sometimes convey a snippet of meaning. However, the computational cost is high and in the end the results obtained without n-grams is sufficient. The most important things are the quality of the data to begin with, and setting up the parameters of the algorithm.

The other approach would be to retain all the words i.e. not remove stopwords, and use techniques from natural language processing (NLP) to cluster the corpus. This would be especially useful when classifying the blogs using the categories drawn form the existing literature. Here, the blogs are grouped more thematically, and it could be argued that the meaning of the words (through word *choice*) is much more important when trying to decide if a blog is 'soapboxing' or expressing a professional opinion in a calm and thoughtful manner. However, in spite of all the hype that exists around language and artificial intelligence, NLP is still based on pattern recognition. Computers do not understand language. What they *can* do is, for example, label 'but' as a conjunction, preposition, adverb or noun depending on its position in the sentence and the words that precede it and come after. In this way, some 'rules' can be devised with the help of the researcher, and work on classification begin.

Behind all the blog posts is a community of people. A lot of them seem to be on Twitter, which they use to promote their own and each-others' blogs, and to have extended discussions about points raised in blogs. Combining Twitter data with blog owners (not an easy task as Twitter and blog author names rarely match) would provide an enriched data set that could be represented as a network of connections. Given that blog posts and tweets also include a timestamp, it would also be possible to track the diffusion of information through the Edu-blog- and Edu-Twitter-sphere which would provide some additional data to see *when* topics became 'hot' in the community, and what prompted the discussion to begin with. This would also reveal the 'influencers' in the community. Comments left below blog posts would also be useful here, as the comments - when not left anonymously - are left under a name or pseudonym which can be linked with a blogger or tweeter.

The blog sites and posts themselves also contain additional data in the form of blog *titles*, and tags devised and attached by the authors. These tags are known as a 'folksonomy', a term fist coined by Thomas Vander Wal in 2004[13]. Tags are particularly useful, as where they exist they are already indicating a class or topic in which they should be included. Multiple tags would suggest multiple topics, which is the strength of topic

[13] https://en.wikipedia.org/wiki/Thomas_Vander_Wal

modelling (groups are not exclusive), and should be given additional weight within the model (Trant, 2009). Indeed, some bloggers include 'tag clouds' as a way of helping the reader navigate through a URL with a substantial number of posts.

Finally, it is worth mentioning again the biases that are inherent in this research. All but the final one have been drawn from Agrawal et al. (2018) and sum up very well the drawbacks and challenges of any research of this kind.

The **sampling** of the data used to develop the methodology may not produce the same results when used on new sets of data, even when the data has been gathered from the same domain.

Learner bias: there are many algorithms available to test on data. This methodology has tested several, but cannot test them all, therefore it is possible that a better classifier could be constructed.

Evaluation bias: While the evaluation of the classifier is based on the accuracy score over test data, the evaluation of the topic model, and the *overall* evaluation of the results is based on the experience and expertise of the researcher.

Order bias: With topic modelling using LDA, each time the algorithm is run the results may me slightly different due to the fact that the data samples are picked at random by the algorithm. Also, changing the order of the data changes the output.

Researcher bias: As mentioned before, the researcher has brought her knowledge, experience, and 'positionality' to this research. It is entirely possible that someone else from the world of Education and Computer Science would take an entirely different route to answer the research questions, and evaluate the results.

6.9 Final Reflections and Original Contributions

This study has shown how important it is for research to be interdisciplinary, especially with regard to Computer Science. In many ways, this is the essence of Web Science, which recognises the importance of the Social Sciences, not least of all because so much data is now available generated by, and from, the Web, which cannot be analysed at scale. What is apparent is that while a Social Scientist does not have to be a Computer Scientist (or vice versa) some critical understanding of how data is transformed, and how algorithms work (and should not be used 'out of the box') is important. Indeed, books are now being published, for example 'Bit By Bit' by Salganik (2018), to help Social Science researchers analyse 'big data'. While software programs such as NVivo[14] have been developed to assist qualitative researchers, the tools it offers computationally

[14] https://www.qsrinternational.com/nvivo/what-is-nvivo

limited. Analysing text data at scale is best performed using code such as Python, and computational libraries written for Python.

The skills necessary to deal with data from the perspective of research originating from the Social Sciences include how to harvest and clean text data; an understanding of how algorithms transform text data to strings of numbers and how those numbers are used to build, for example, distance models (and what that means for data analysis); and the benefits and limitations for the various analytical models. Even harvesting data efficiently from the Web requires coding skills and a way of thinking through the problem from the point of view of a Web/Computer Scientist. This approach has much in common with the emerging field of 'Quantitative Ethnography' (Schaffer, 2017). The first conference focusing on this field was held in 2019 (Eagan et al., 2019) and included papers on topic modelling (Bakharia, 2019) and the use of classifiers (Lee et al., 2019). The focus of the conference was a recognition that 'data' has moved beyond that which was collected via interviews and surveys, and is now widely available, in 'big data' quantities, but requires different skills (from the Social Science community, at least) to access and analyse.

The tools and techniques used to answer the research questions posed by *this* thesis is one of the two original contributions of this research.

The second original contribution is a detailed oversight of the online discourse of a large, professional community. This has hitherto only been carried out in a modest way i.e. with small sample sizes (and here I refer to the *content* of social media posts) and largely hand-coded. The themes arising from existing research have been used to categorise blog posts going back to 2004; topic modelling has revealed a wealth of 'new' topics including the impact of issues arising from a variety of sources. The direct impact of Government policy is also reflected. Fifty-three topics were labelled, covering everything from 'Art' to 'Writing', their ebb and flow through the community evident year-on-year.

As a direct result of an engagement with tools from Computer Science, this research has shown the Edu-community to broad, vibrant, opinionated and concerned about the teaching profession. This is a community through which almost every child will pass on their way to taking up their positions in society. That most children do this successfully is testament to the hard work that is put in by the community, and their willingness to adapt to changes imposed on them from above. In Peal (2015), Andrew Old made the case that"the old consensus in education no longer exists on social media" (Peal, 2015, Ch.5, p.64) and that challenges to every orthodoxy exist online. While this research has done no more than label the topics that exist in the blog posts collected, I have no doubt that within the topic 'teaching practice', for example, will exist opposing views as to the 'best' method. This is to be celebrated, not least of all because it should prompt practitioners to continue to question what they are doing, and look for *evidence* to support their views and methods. This only strengthens their position, and the need for teacher voice to continue to be heard in government.

Appendix A

Stop Words

A.1 scikit-Learn Stop Word Set

a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond bill both bottom but by call can cannot cant co con could couldnt cry de describe detail do done down due during each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fifty fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred i ie if in inc indeed interest into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mill mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps please put rather re same see seem seemed seeming seems serious several she should show side since sincere six sixty so some somehow someone something sometime sometimes somewhere still such system take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thick thin third this those though three through throughout thru thus to together too top toward towards twelve twenty two un under until up upon us very via was we well were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves

A.2 NLTK Stop Word Set

a about above after again against all am an and any are as at be because been before being below between both but by can did do does doing don down during each few for from further had has have having he her here hers herself him himself his how i if in into is it its itself just me more most my myself no nor not now of off on once only or other our ours ourselves out over own s same she should so some such t than that the their theirs them themselves then there these they this those through to too under until up very was we were what when where which while who whom why will with you your yours yourself yourselves

A.3 Bespoke Stop Words

a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond bill both bottom but by call can cannot cant co con could couldnt cry de describe detail do done down due during each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fifty fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred i ie if in inc indeed interest into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mill mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps please put rather re same see seem seemed seeming seems serious several she should show side since sincere six sixty so some somehow someone something sometime sometimes somewhere still such system take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thick thin third this those though three through throughout thru thus to together too top toward towards twelve twenty two un under until up upon us very via was we well were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves student students school teach teacher teachers teaching time year work use like make need think question lesson lessons

Appendix B

Topic Modelling Visualisations

B.1 Topic Modelling

Visualising data that is large, complex and multi-dimensional presents challenges. The data must be summarised in such a way that the important information is retained, whilst the less important data is either discarded or combined with other data before being summarised. A high-dimensional data set needs to be reduced to a low dimensional graph so that it can be plotted on a number line.

Topic modelling starts with a word count matrix, which is use to calculate the probability that word *a* is more likely to represent a topic being discussed in document *x* than word *b*, and so on. It is then a simple matter to produce a table listing the top 5 or 10 words in each of a specified number of topics. In effect, the LDA algorithm is grouping documents according to the *dominant topic*. It would be interesting and useful to see a visual representation of the *documents* as opposed to a table of words. There are different algorithms that can achieve this; one is presented here because it is both informative, efficient and generally gives an accurate representation within.

The dimensionality-reduction algorithm t-SNE (t distribution-Stochastic neighbourhood embedding) is applied (or *fitted*) to the LDA matrix. tSNE uses the local relationship between points to create a low-dimensional map. It is non-deterministic, which means that the graph will vary slightly every time the algorithm is run. The process is explained very well here[1].

The topic modelling data presented here shows each topic (for each year) represented in a different colour, with the top 5 words added for additional information. The colours are not always grouped together: in 2011, for example, it is possible to see that there are dark blue dots clustered around the words 'ict, confer(ence), technolog(y), present,

[1] https:www.youtube.comwatch?v=NEaUSP4YerM&t=69s

meet' and below 'tool, moodl(e), simpl(e), video, inform'. Topic modelling is also non-deterministic, and even if we were to assume the tSNE model is accurate, we cannot assume the same of LDA. Nevertheless, the plot presents a useful and informative view of the underlying data once it has been processed, and the distribution of colours suggests that the underlying topic model is reasonable calculating 20 topics.

Because the data points are based on a distance model, they may be plotted in -*x,y* co-ordinates.

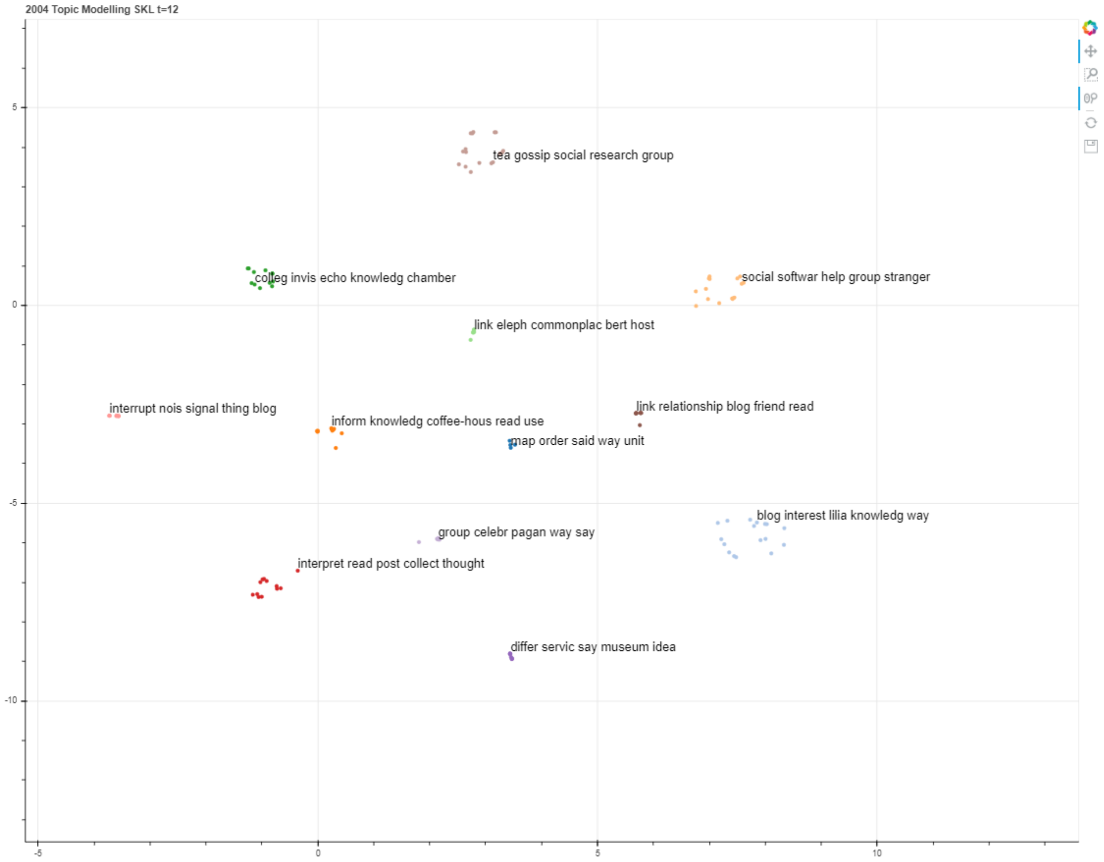


Figure B.1: Representation of Topic Clusters Using tSNE (2004)

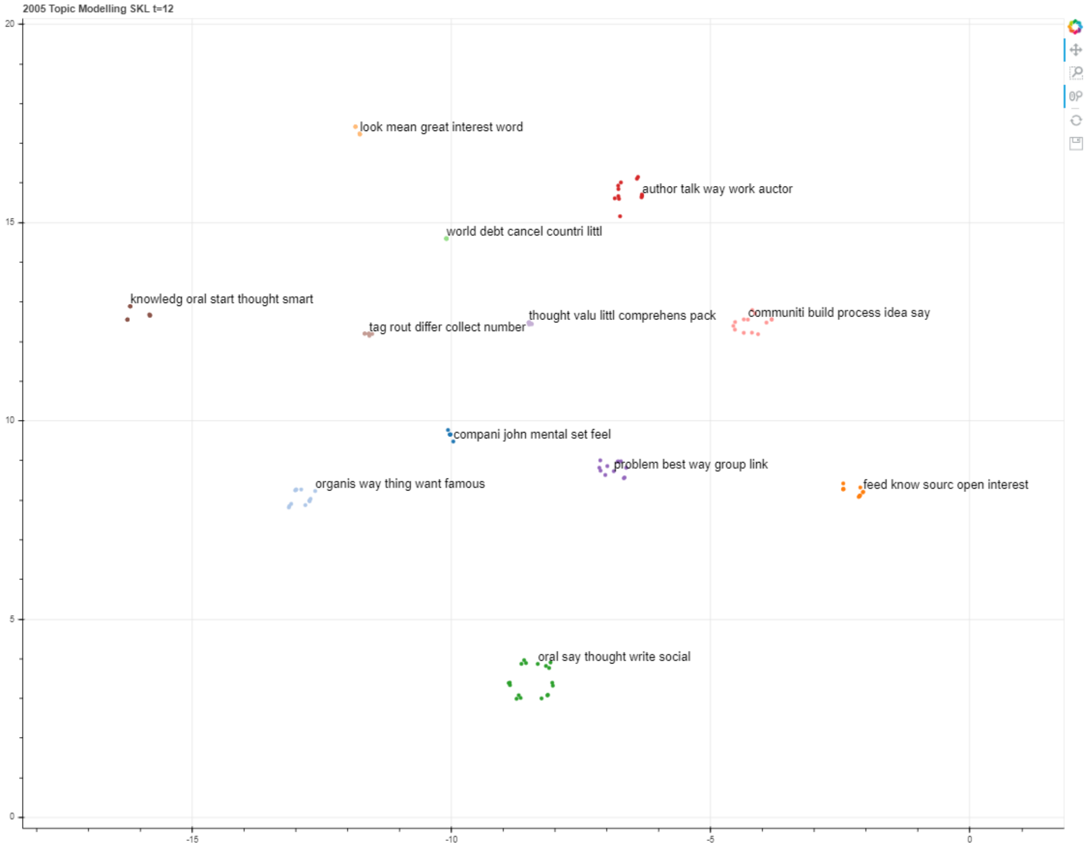


Figure B.2: Representation of Topic Clusters Using tSNE (2005)

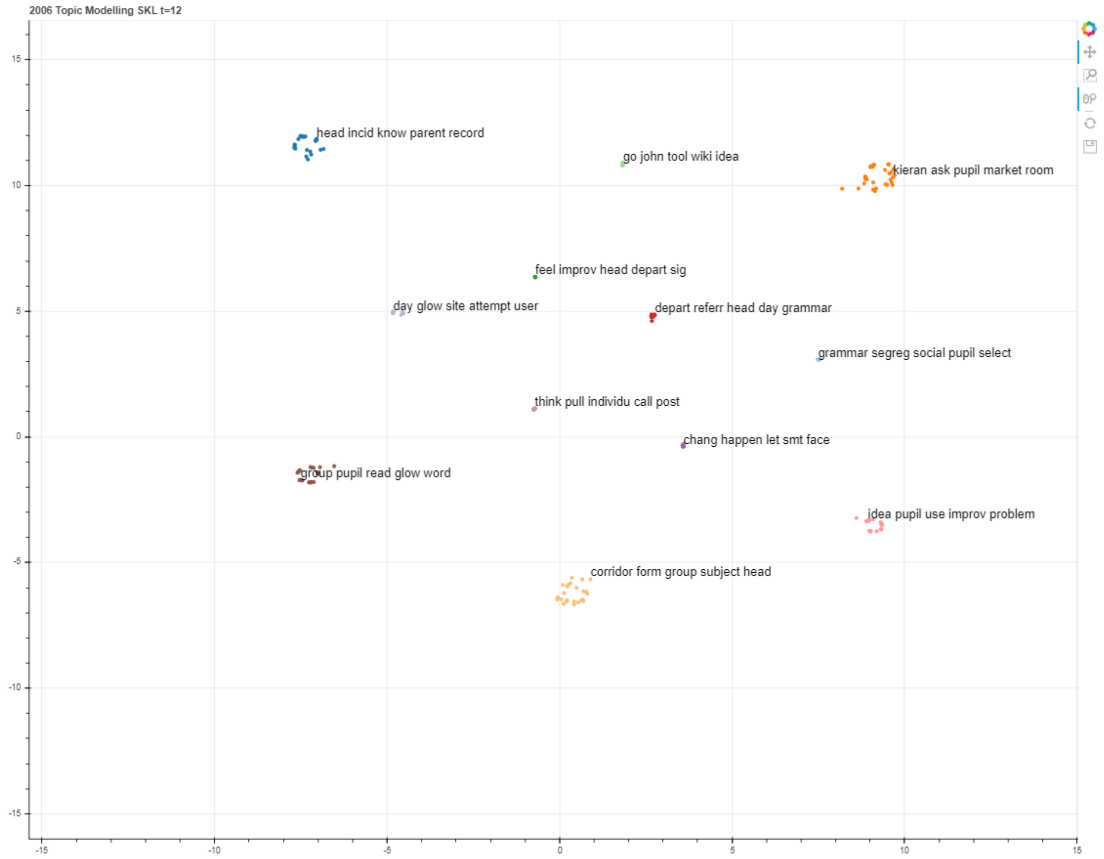


Figure B.3: Representation of Topic Clusters Using tSNE (2006)

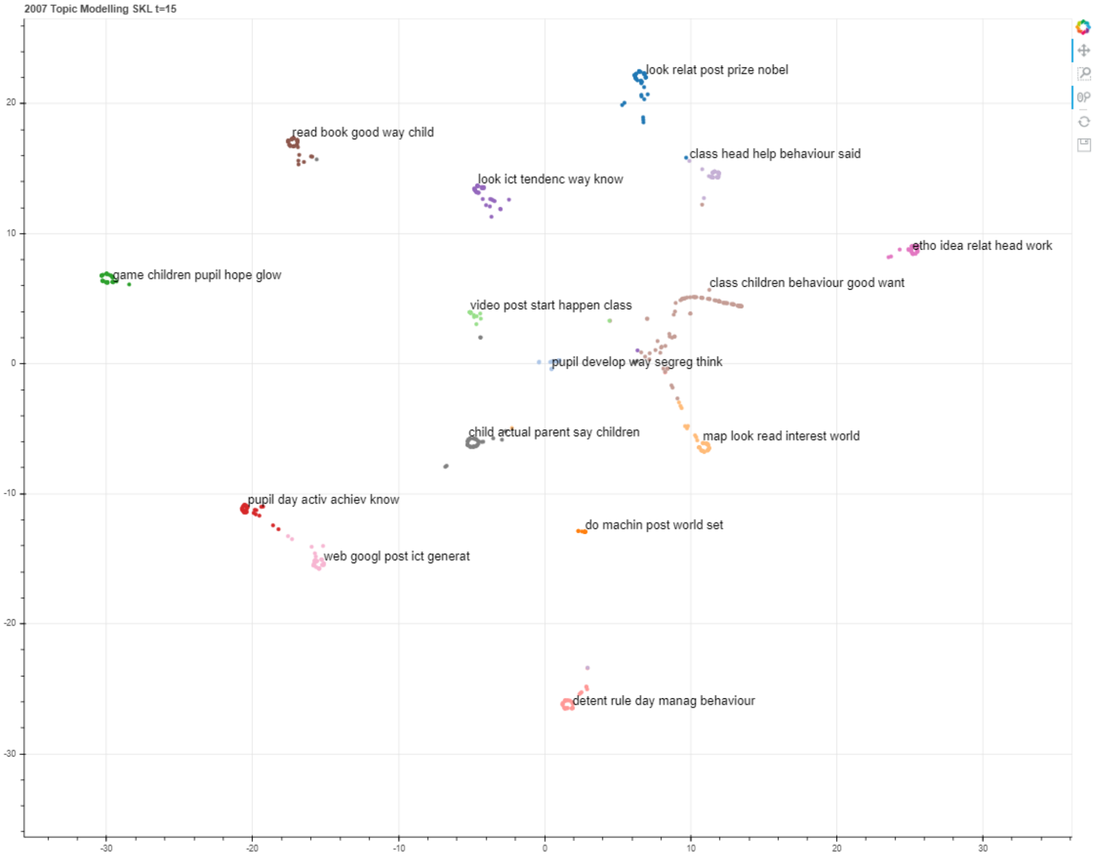


Figure B.4: Representation of Topic Clusters Using tSNE (2007)

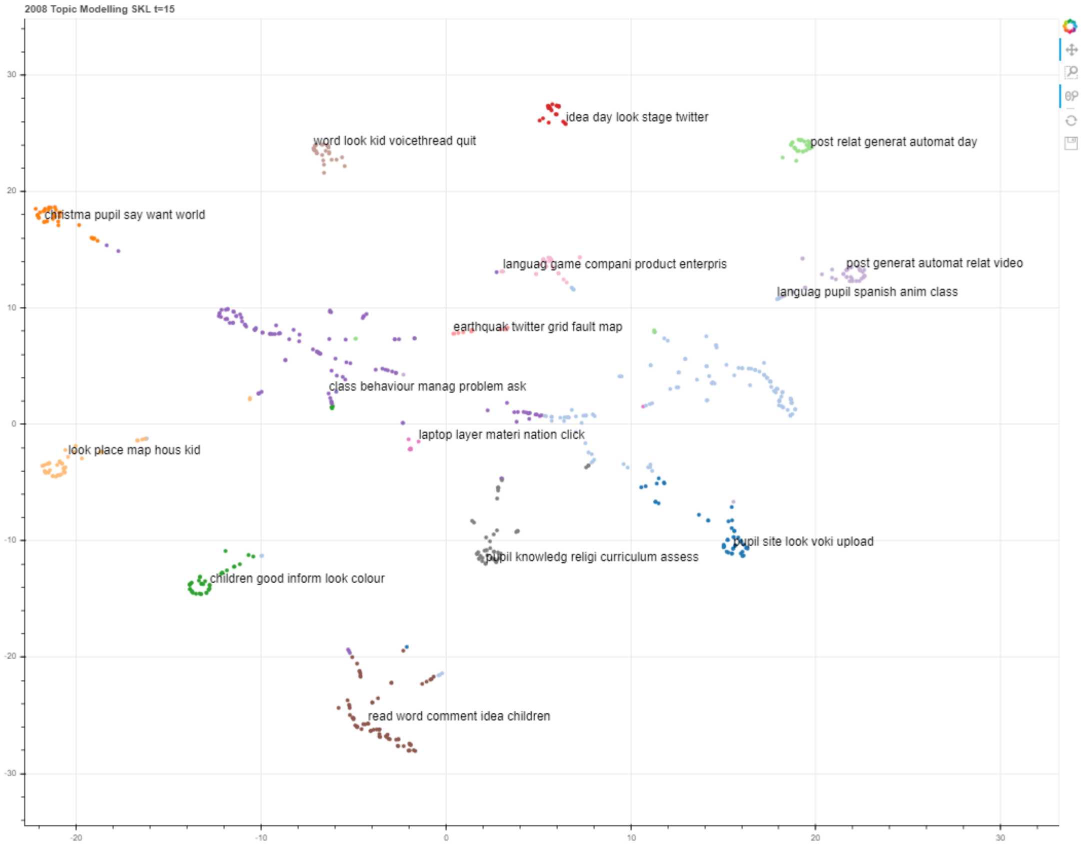


Figure B.5: Representation of Topic Clusters Using tSNE (2008)

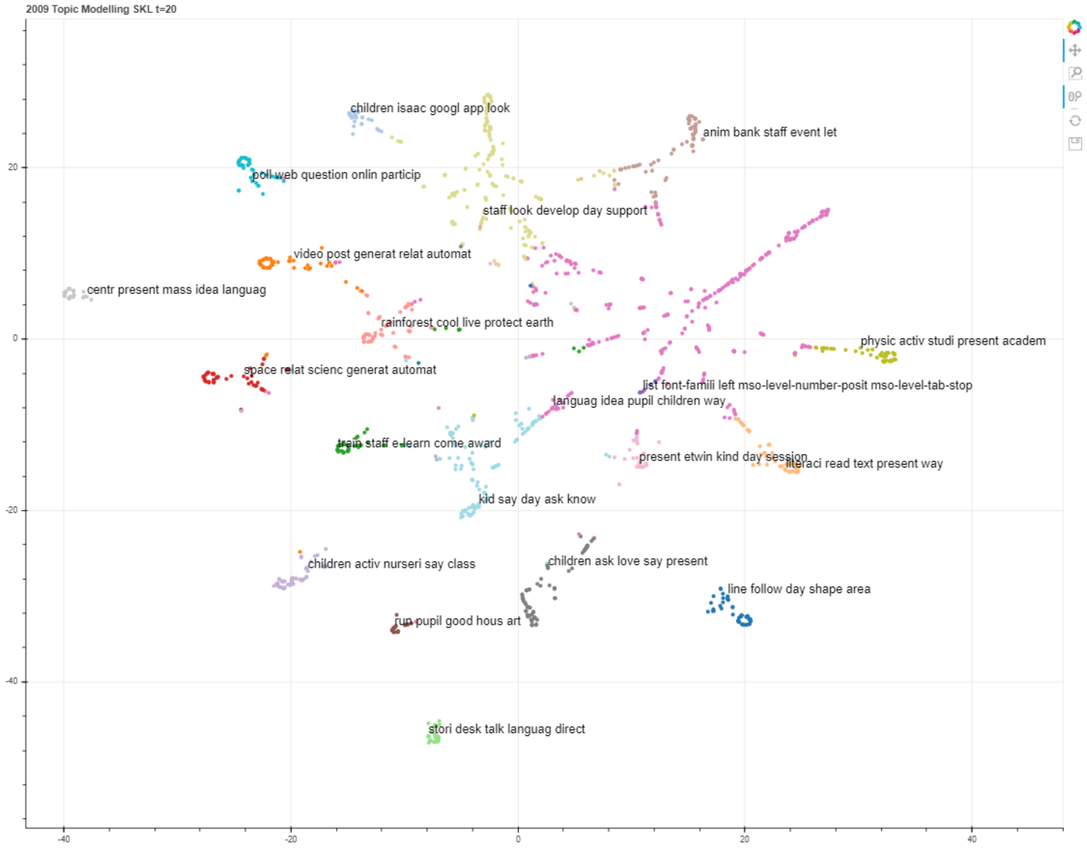


Figure B.6: Representation of Topic Clusters Using tSNE (2009)

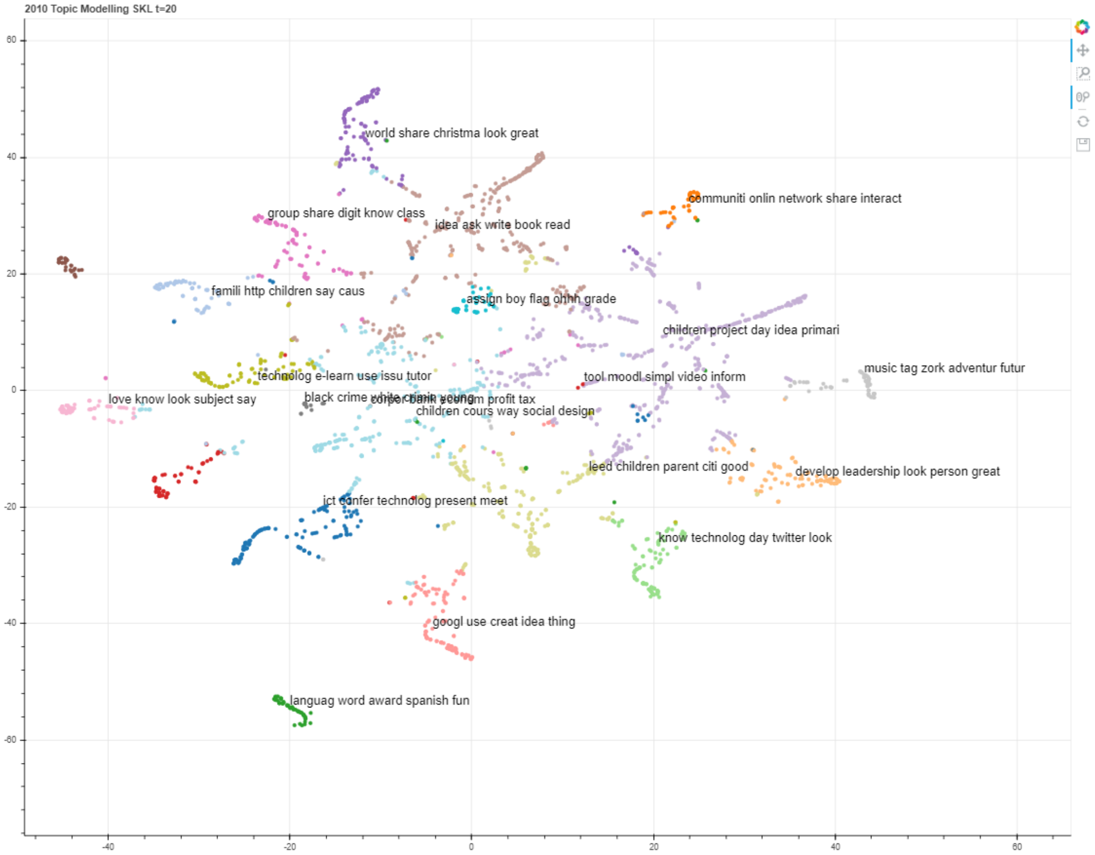


Figure B.7: Representation of Topic Clusters Using tSNE (2010)

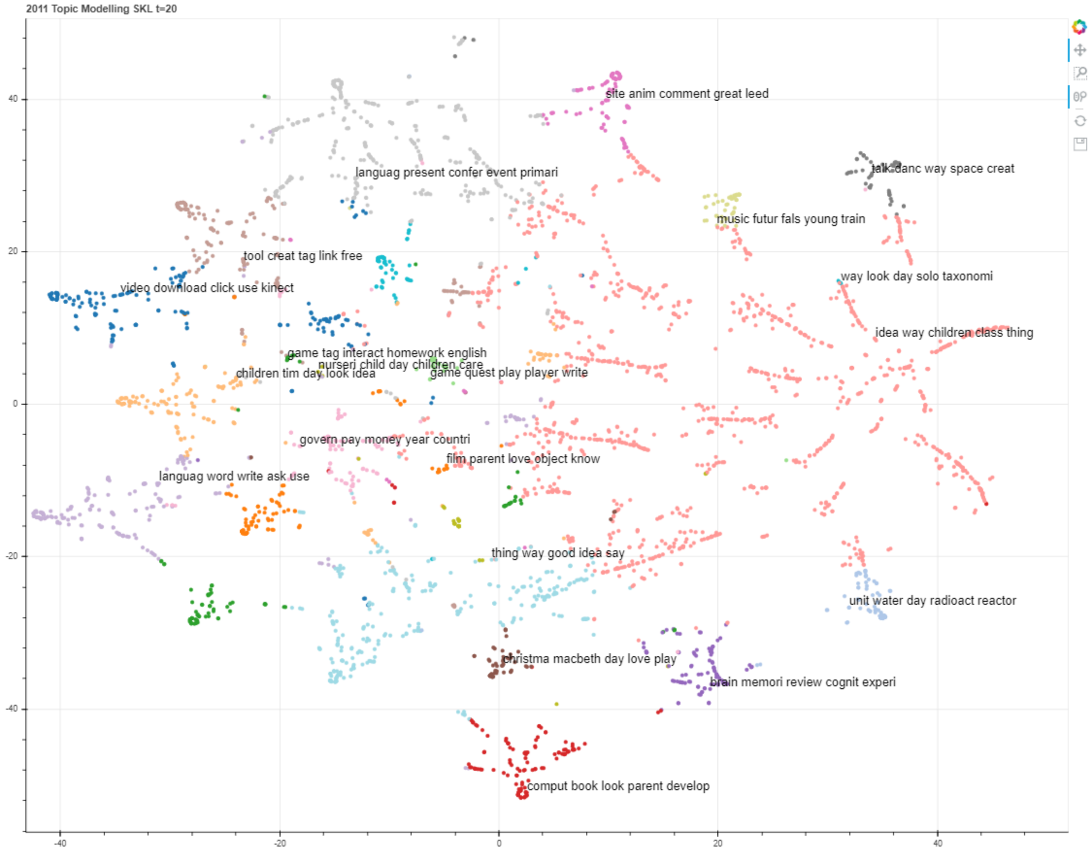


Figure B.8: Representation of Topic Clusters Using tSNE (2011)

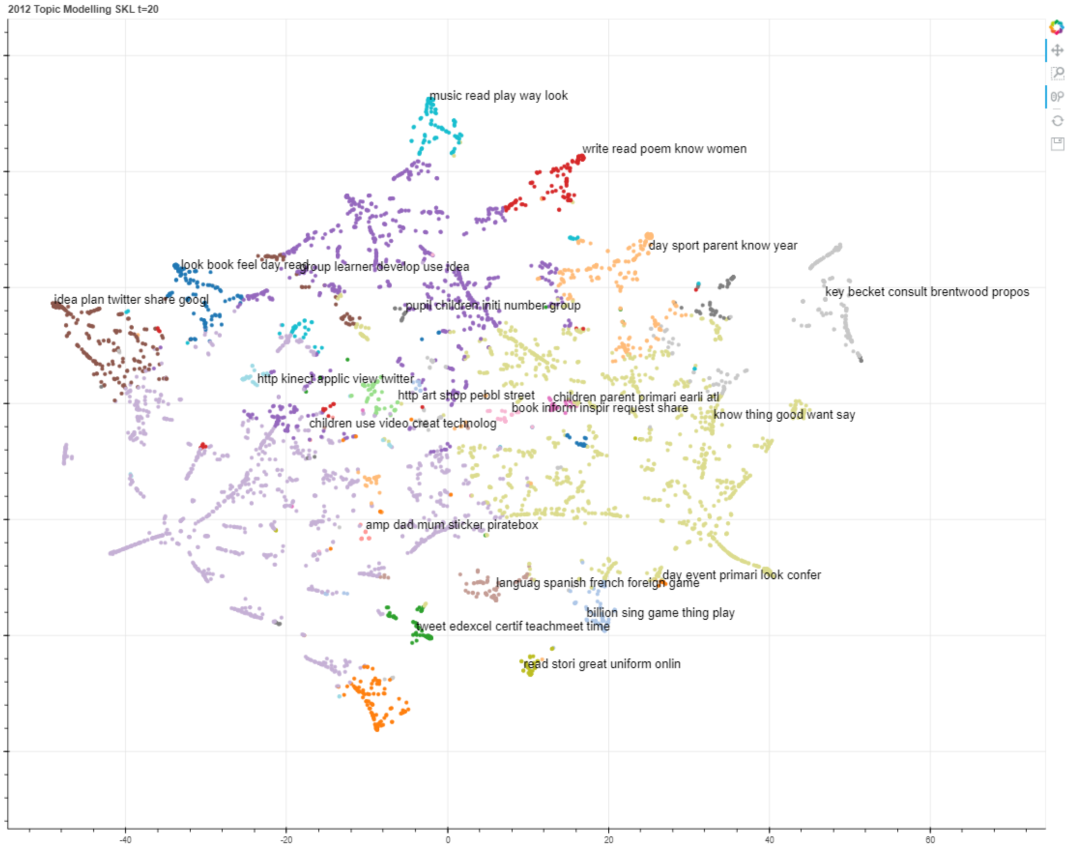


Figure B.9: Representation of Topic Clusters Using tSNE (2012)

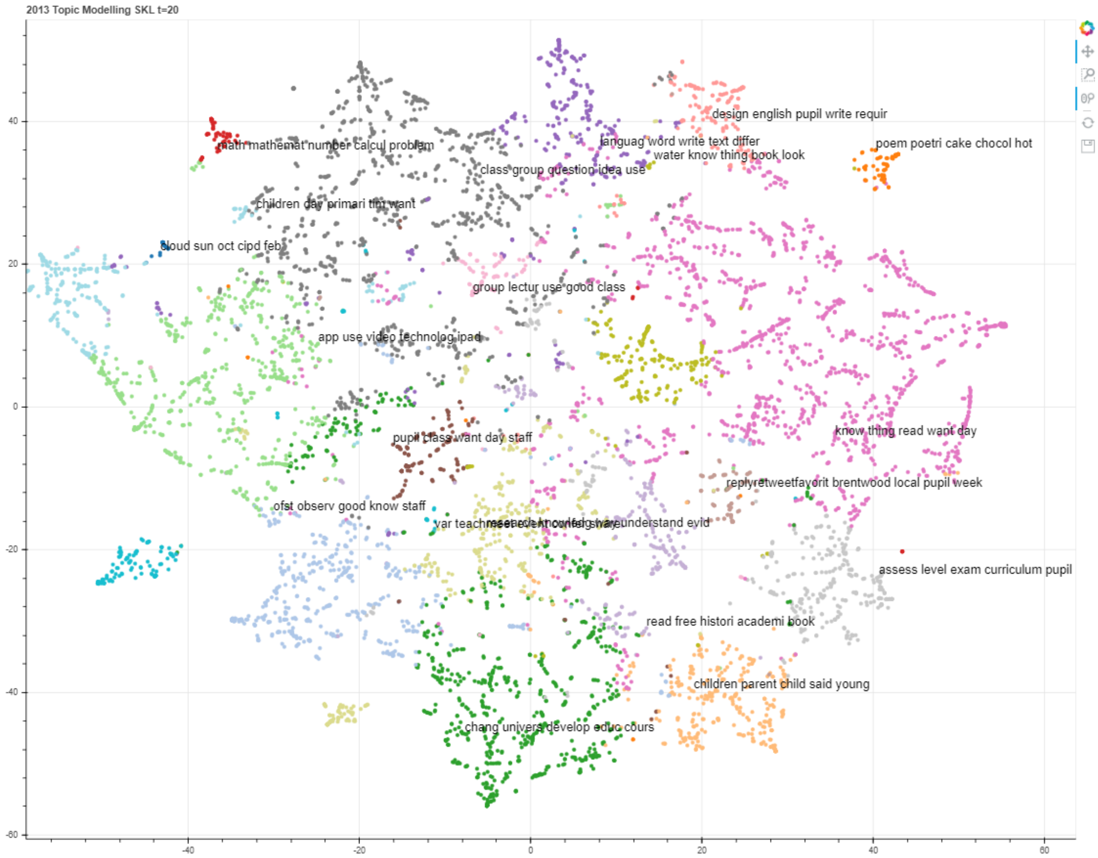


Figure B.10: Representation of Topic Clusters Using tSNE (2013)

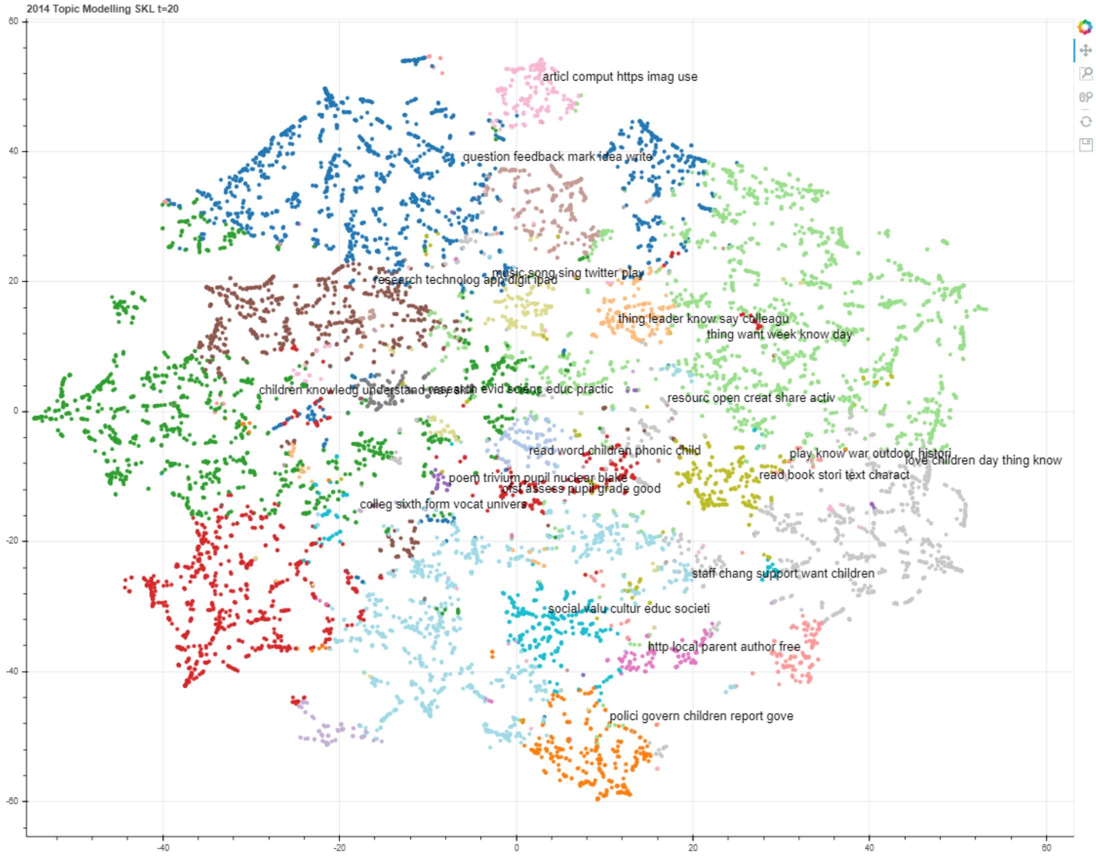


Figure B.11: Representation of Topic Clusters Using tSNE (2014)

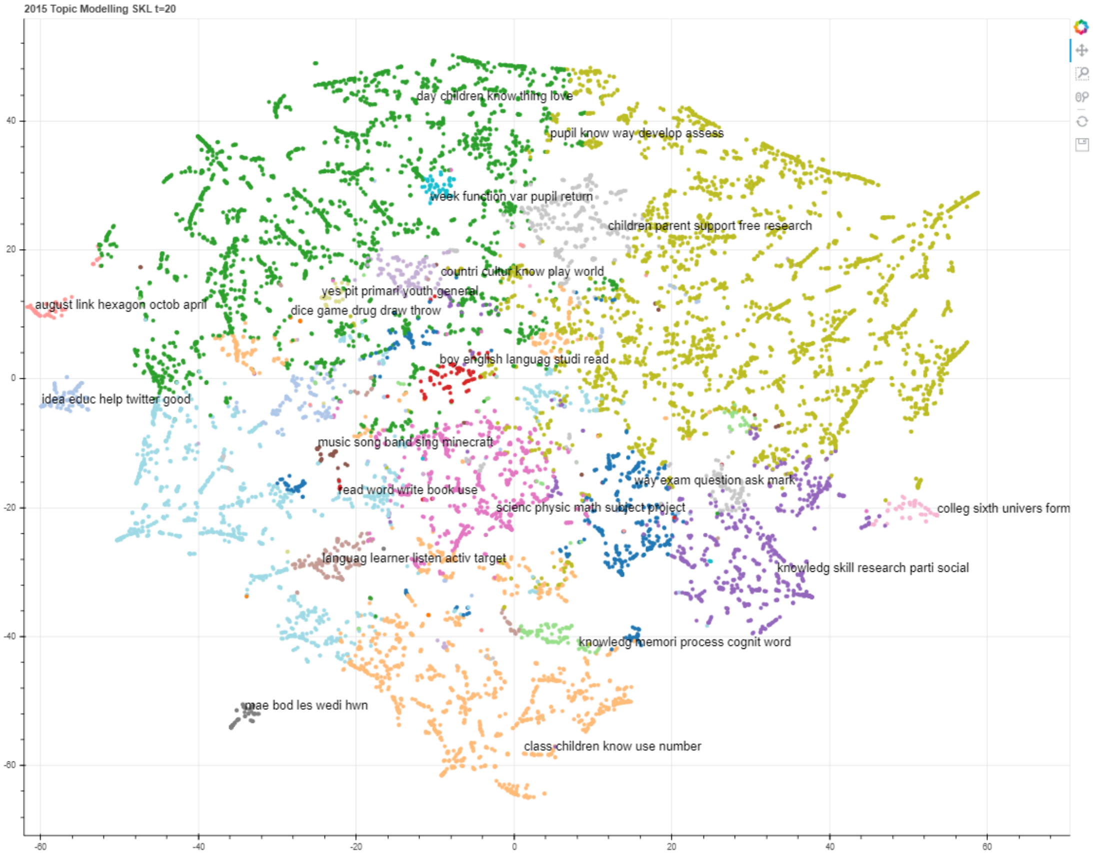


Figure B.12: Representation of Topic Clusters Using tSNE (2015)

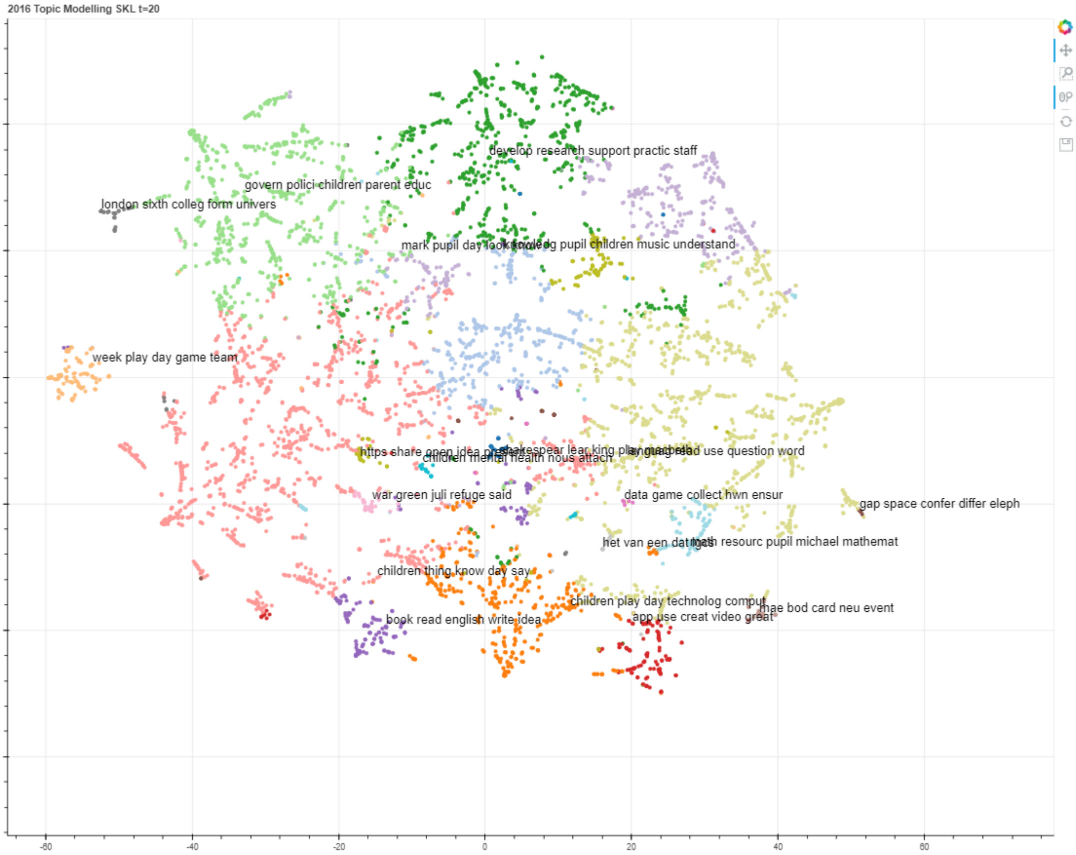


Figure B.13: Representation of Topic Clusters Using tSNE (2016)

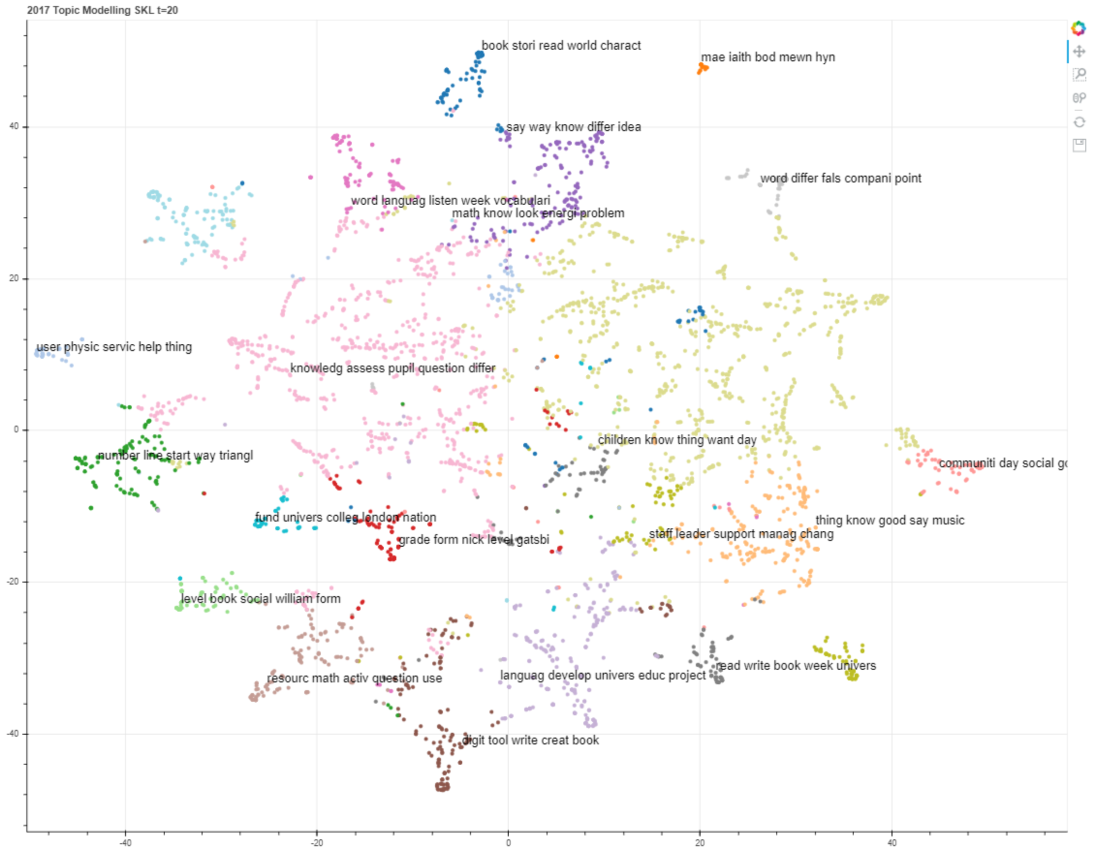


Figure B.14: Representation of Topic Clusters Using tSNE (2017)

Appendix C

Topic Modelling Results Tables

2004													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	**TK**
0	**8**	**Criminal, crime, poem**	said	order	feel	develop	littl	cure	phone	crime	crimin	unit	CRI
1	**12**	**Knowledge, (challenge)**	knowledg	lilia	manag	way	blog	ton	person	know	decis	differ	KNO
2	**11**	**Inform, knowledge**	inform	knowledg	coffee-hous	person	differ	use	number	type	keep	rule	KNO
3	**14**	**Social, software, map**	social	softwar	map	help	group	way	read	thing	call	post	SOC
4	**10**	**College, echo chamber**	colleg	invis	echo	chamber	read	knowledg	use	allow	visibl	call	MISC
5	**10**	**Elephant, poetry**	thing	eleph	commonplac	bert	host	link	book	ingrid	happen	comment	MISC
6	**9**	**Interpret, (philosophy), text**	interpret	text	german	hermeneut	author	understand	thought	way	live	knowledg	PHIL
7	**10**	**Interrupt, noise, signal**	interrupt	thing	nois	signal	say	interest	ton	task	collect	human	MISC
8	**10**	**(Museum), network, metaphor**	network	human	idea	differ	inform	littl	blog	tool	metaphor	hard	SOC
9	**7**	**Pagan, celebration**	group	celebr	pagan	small	post	church	anil	crowd	way	soon	MISC
10	**10**	**Stranger, familiar, friend**	stranger	friend	familiar	relationship	interest	link	blog	thing	go	good	SOC
11	**15**	**Tea, gossip, social**	tea	gossip	group	social	interest	say	good	research	make	point	SOC

2005													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**7**	**Company, mental, (trackback)**	compani	john	mental	set	feel	media	think	comment	communiti	etc.	MISC
1	**12**	**Organise, famous, autoblogg(er, ing)**	organis	way	thing	want	famous	autoblogg	tag	tri	interest	organ	MISC
2	**11**	**Feed, source**	feed	know	sourc	open	interest	differ	say	group	link	want	MISC
3	**9**	**Look, cricket**	look	mean	great	interest	word	network	social	know	ontolog	organis	SOC
4	**18**	**Oral, say, thought**	oral	say	thought	write	social	know	make	studi	group	read	MISC
5	**6**	**World debt**	world	debt	cancel	countri	littl	round	brown	that.	poor	general	BAN
6	**13**	**Author, talk**	author	talk	way	work	auctor	want	have	anjo	grow	problem	MISC
7	**13**	**Community, build, (coffee)**	communiti	build	process	idea	say	love	interest	thing	week	paper	SOC
8	**12**	**Problem, best, practice**	problem	best	way	group	link	practic	tri	interest	research	suggest	TP
9	**10**	**Comprehensive, pack, (airport)**	thought	valu	littl	comprehens	pack	provid	contribut	actual	end	process	MISC
10	**12**	**Knowledge, philosophy, Homer**	knowledg	oral	start	thought	smart	homer	approach	say	train	think	PHIL
11	**8**	**Route, collect, hierarchy**	tag	rout	differ	collect	number	type	hierarchi	know	think	item	MISC

2006												
Topic Num	**Num Documents**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**26**	**Record, incident, (behaviour)**	head	incid	know	parent	record	detent	want	way	have	children	BEH
1	**13**	**Grammar schools**	grammar	segreg	social	pupil	select	lea	schools.	secondari	local	abil	GRA
2	**30**	**Behaviour**	kieran	ask	pupil	market	room	tri	good	black	class	head	BEH
3	**28**	**Corridor, form, pshe**	corridor	form	group	subject	head	day	actual	know	depart	tell	MISC
4	**5**	**Inprov(e, ing), head, depart(ment)**	feel	improv	head	depart	sig	suggest	point	look	allow	staff.	MISC
5	**15**	**Tool, wiki, ICT**	go	john	tool	wiki	idea	kid	blog	whilst	imagin	ict	TECH
6	**18**	**Depart(ment), referr(ing), grammar**	depart	referr	head	day	grammar	let	ask	gemma	nevill	crosland	MISC
7	**23**	**Idea, assess, (Gardner)**	idea	pupil	use	improv	problem	educ	assess	parent	lie	abil	ASS
8	**13**	**SMT, behaviour**	chang	happen	let	smt	face	pupil	result	behaviour	believ	act	BEH
9	**19**	**Glow, site**	day	glow	site	attempt	user	chalk	child	access	sharepoint	start	GLOW
10	**25**	**Group, read, Glow**	group	pupil	read	glow	word	class	know	come	set	staff	GLOW
11	**14**	**Think, individual**	think	pull	individu	call	post	anjo	group	good	local	say	MISC

2007												
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**26**	**Democrat(ic), language, culture**	way	idea	democrat	develop	pupil	languag	say	cultur	good	look	CUL
1	**44**	**etwin, post, game**	post	world	project	game	generat	curriculum	play	etwin	good	start	MISC
2	**36**	**Read, discover**	pupil	read	let	children	relat	have	want	look	great	discov	READ
3	**18**	**Map, look, read**	map	mice	bank	press	build	cave	game	click	support	look	MAP
4	**84**	**(Software), read, tool**	pupil	read	look	tri	develop	help	provid	start	resourc	tool	READ
5	**44**	**Post, web**	post	web	read	relat	generat	day	automat	star	idea	come	MISC
6	**40**	**ICT, technology, chang(e, ing)**	ict	chang	technolog	differ	children	pupil	social	comput	say	curriculum	TECH
7	**16**	**Behaviour**	etho	site	flickr	upload	click	origin	post	share	respons	develop	BEH
8	**49**	**Journey, travel**	way	long	day	start	good	tri	old	walk	make	world	TRA
9	**34**	**Duty, break**	interest	look	day	duti	year	thing	break	pupil	work	data	MISC
10	**69**	**Behaviour**	set	manag	head	rule	detent	behaviour	good	look	day	thing	BEH
11	**68**	**Behaviour**	class	children	behaviour	actual	child	classroom	parent	ask	problem	know	BEH

2008													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**41**	**Park, laptop**	pupil	park	origin	want	laptop	right	good	world	way	layer	MISC
1	**29**	**Upload, wiki**	pupil	upload	share	comment	wiki	place	hope	look	set	thought	TECH
2	**36**	**Language, company, site, (elearn)**	languag	compani	good	site	look	group	rey	game	glow	start	TECH
3	**100**	**Language, Spanish, primary**	languag	pupil	spanish	idea	primari	class	video	children	day	word	MFL
4	**32**	**Twitter, (duty), tweet**	twitter	idea	day	boy	present	thing	want	blog	tweet	start	SOC
5	**37**	**Map, house, mental**	map	hous	start	inform	look	draw	mental	thing	say	lot	MAP
6	**27**	**Place, map, house**	page	grid	add	quit	map	way	call	hand	click	request	MAP
7	**31**	**Read, settlement, (India)**	read	settlement	pupil	page	word	function	inform	set	talk	share	READ
8	**59**	**Word, look, Spanish (Voki, animoto)**	anim	word	look	spanish	unit	music	upload	play	site	pupil	MFL
9	**37**	**Religi(on), select, grammar**	pupil	religi	social	select	children	secondari	admiss	grammar	comprehens	test	GRA
10	**36**	**Look, comment, (Catalan)**	look	comment	ask	day	say	origin	week	upload	interest	farm	TRA
11	**110**	**Behaviour**	class	behaviour	children	manag	parent	problem	say	actual	head	want	BEH
12	**31**	**Volcano, language, assess**	languag	session	assess	day	volcano	pupil	primari	activ	look	present	LANG
13	**26**	**Site, (kame, cinco de mayo, Geography?)**	site	day	idea	support	book	good	matild	know	pupil	comment	GEOG
14	**38**	**Knowledge, curriculum, (structure)**	knowledg	curriculum	differ	subject	point	base	good	young	everyday	practic	CR

2009													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**52**	**(Detention), shape, line, angle**	line	follow	day	shape	area	detent	simpl	reflect	do	head	MATH
1	**33**	**Isaac, (dub), google**	children	isaac	googl	app	look	said	year	number	recent	song	TECH
2	**52**	**Video, post, prize, (NASA)**	video	post	generat	relat	automat	prize	physic	thou	idea	word	MISC
3	**56**	**Literacy, read, text**	literaci	read	text	present	way	develop	word	day	twitter	poster	READ
4	**50**	**E-learn, train, staff**	train	staff	e-learn	come	award	custom	employe	way	knowledg	product	TSD
5	**31**	**Story, desk, talk**	stori	desk	talk	languag	direct	door	crash	love	airbag	varna	LANG
6	**53**	**Space, science, world**	space	relat	scienc	generat	automat	good	world	way	truth	post	SCI
7	**62**	**Rainforest, cool, earth (geography?)**	rainforest	cool	live	protect	earth	punish	self-esteem	bad	talk	good	GEOG
8	**10**	**MSO (web design)**	list	font-famili	left	mso-level-tab-stop	mso-level-number-posit	time	mso-level-text	text-ind	roman	bullet	TECH
9	**56**	**Nursery, active**	children	activ	nurseri	say	class	read	present	good	chang	stori	EY
10	**25**	**(Ladder), politician, (housing)**	run	pupil	good	hous	art	politician	bank	subject	idea	imag	POL
11	**73**	**Bank, staff, (banking?)**	anim	bank	staff	event	let	world	work	game	ask	use	BAN
12	**334**	**Language, idea, class**	languag	idea	pupil	children	way	class	look	use	creat	link	LANG
13	**37**	**Present, session**	present	etwin	kind	day	session	lisibo	team	project	view	interest	EVENT
14	**74**	**Children, ask, (schema)**	children	ask	love	say	present	cours	child	primari	look	number	PRI
15	**33**	**Centre, mass, (fuel, gallon)**	centr	present	mass	idea	languag	say	day	save	primari	start	MISC
16	**51**	**Spanish (poem); physic(al), study**	physic	activ	studi	present	academ	bodi	mark	mientra	idea	haya	MFL
17	**173**	**Staff, network, (mobile)**	staff	look	develop	day	support	group	network	share	ict	technolog	SOC
18	**44**	**Poll, online, participation**	poll	web	question	onlin	particip	way	like	beat	idea	entri	POL
19	**115**	**Kid, play, say**	kid	say	day	ask	know	pupil	thing	children	go	play	MISC

2010												
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**160**	**ICT, conference, technology**	ict	confer	technolog	present	meet	support	microsoft	develop	free	share	TECH
1	**90**	**Family, poverty (violence)**	famili	http	children	say	caus	reason	come	women	univers	poverti	POL
2	**52**	**Community, online, network**	communiti	onlin	network	share	interact	member	develop	social	grant	support	SOC
3	**111**	**Develop, leadership**	develop	leadership	look	person	great	talk	work	adult	book	inform	TSD
4	**64**	**Language, word, Spanish**	languag	word	award	spanish	fun	sound	music	phonic	languages.	way	MFL
5	**109**	**Know, technology, twitter (ocr)**	know	technolog	day	twitter	look	follow	go	thought	ocr	expect	TECH
6	**80**	**Tool, moodle (Technology)**	tool	moodl	simpl	video	inform	improv	look	use	data	live	TECH
7	**117**	**Google, technology**	googl	use	creat	idea	thing	site	share	technolog	app	way	TECH
8	**126**	**Christmas**	world	share	christma	look	great	thing	play	creat	use	video	CHR
9	**481**	**Children, project primary**	children	project	day	idea	primari	languag	present	way	develop	pupil	PRO
10	**37**	**Corporation, bank**	corpor	bank	econom	profit	tax	world	countri	compani	labour	capit	BAN
11	**320**	**Idea, book, read, (schappi)**	idea	ask	write	book	read	comment	know	look	start	answer	READ
12	**119**	**Group, share, digital (animoto)**	group	share	digit	know	class	leader	tri	hope	plan	help	TECH
13	**67**	**Love, maths, (equation)**	love	know	look	subject	say	math	women	feel	team	start	MATH
14	**21**	**Crime, justice**	black	crime	white	crimin	young	depart	english	justic	crime.	visit	CRI
15	**70**	**Music, Zork, game**	music	tag	zork	adventur	futur	game	english	read	stori	place	MUS
16	**108**	**Technology, e-learn, (jisc)**	technolog	e-learn	use	issu	tutor	cours	provid	experi	onlin	resourc	TECH
17	**251**	**Leeds, city, (closure)**	leed	children	parent	citi	good	year	local	day	idea	week	LEEDS
18	**45**	**Assign(ment), boy, flag**	assign	boy	flag	ohhh	grade	task	research	que	freedom	avatar	MISC
19	**251**	**Children, design, course**	children	cours	way	social	design	develop	differ	chang	research	experi	MISC

2011													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**184**	**Video, download, kinect**	video	download	click	use	kinect	week	link	code	read	post	TECH
1	**66**	**Radioactive, reactor, water, (Fukishima)**	unit	water	day	radioact	reactor	power	function	nuclear	thing	sectionid	SCI
2	**93**	**Gang, film, (Mercutio)**	film	parent	love	object	know	say	point	translat	class	gang	PLAY
3	**167**	**Children, time, day**	children	tim	day	look	idea	way	great	child	want	tri	MISC
4	**94**	**Game, homework, interact(ive)**	game	tag	interact	homework	english	quest	text	fiction	play	creat	TECH
5	**33**	**Game, quest, player**	game	quest	play	player	write	reward	task	plan	simpl	red	TECH
6	**119**	**Comput(er), book, women, girl**	comput	book	look	parent	develop	women	boy	girl	use	world	TECH
7	**1221**	**Idea, way, class**	idea	way	children	class	thing	develop	want	know	use	start	MISC
8	**94**	**Brain, memory, dementia**	brain	memori	review	cognit	experi	human	dementia	idea	object	person	MC
9	**189**	**Language, word, (vocabulary)**	languag	word	write	ask	use	english	activ	learner	differ	idea	LANG
10	**54**	**Christmas, Macbeth, play, Shakespeare**	christma	macbeth	day	love	play	scene	act	trade	fair	shakespear	PLAY
11	**156**	**Tool, create, (moodle)**	tool	creat	tag	link	free	use	present	onlin	video	look	TECH
12	**91**	**Site, animation, (Leeds)**	site	anim	comment	great	leed	look	want	way	creat	curriculum	TECH
13	**106**	**Goverm(ment), pay, tax**	govern	pay	money	year	countri	tax	public	compani	million	world	POL
14	**63**	**Talk, dance, space**	talk	danc	way	space	creat	play	point	object	ipad	thought	MISC
15	**339**	**Language, conference, present**	languag	present	confer	event	primari	day	great	share	project	ict	LANG
16	**29**	**Nursery, early**	nurseri	child	day	children	care	knowledg	pencil	earli	hat	start	EY
17	**45**	**Music, future, false**	music	futur	fals	young	train	review	x-none	champion	share	support	MUS
18	**58**	**Taxonomy, solo**	way	look	day	solo	taxonomi	number	idea	univers	christma	best	ST
19	**276**	**Science, (happiness)**	thing	way	good	idea	say	know	read	mean	scienc	point	MISC

2012													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**110**	**Look, book, read**	look	day	book	feel	love	read	great	ask	challeng	want	READ
1	**90**	**Billion, sing, world**	billion	sing	play	thing	world	game	fact	relat	number	choir	MUS
2	**154**	**Day, event, primary, confer(ence)**	day	event	confer	primari	world	categori	hello	social	comment	word	EVENT
3	**200**	**Day, sport, olympic**	day	sport	know	live	year	job	olymp	chang	way	want	SPO
4	**85**	**Edexccel, certificate, iGCSE**	tweet	certif	edexcel	time	igcs	follow	support	star	protect	day	QUAL
5	**45**	**AMP, art**	amp	http	pebbl	art	street	shop	museum	town	station	bedford	MISC
6	**128**	**Poem, poetry, women**	write	read	poem	women	idea	know	say	point	poetri	languag	READ
7	**30**	**Dad, mum (Ada Lovelace, Babbage)**	dad	mum	day	earli	test	sticker	year	fair	nurseri	sister	TECH
8	**713**	**Learner, group, develop, activ(e, ity)**	learner	group	develop	idea	way	use	question	differ	discuss	activ	TP
9	**1046**	**Video, create, technology**	children	use	video	creat	technolog	way	share	app	look	game	TECH
10	**301**	**Idea, plan, Twitter, share, Google**	idea	twitter	plan	share	good	know	tri	week	want	start	TECH
11	**66**	**Language, (MFL)**	languag	spanish	french	foreign	univers	cours	game	german	particip	english	MFL
12	**34**	**Children, parent, carer, ATL**	children	parent	carer	band	atl	meet	earli	data	year	communiti	ATL
13	**27**	**Inform(ation), request, inspir(e, ation)**	inform	book	request	share	inspir	free	kid	releas	sport	nuclear	MISC
14	**73**	**Pupil, init(ially), group**	pupil	initi	children	number	group	look	place	parent	support	water	MISC
15	**249**	**Key, consult(ant, ing), Brentwood, Becket**	key	becket	consult	brentwood	propos	provid	parent	fund	http	free	ACD
16	**48**	**Read, uniform, MOOC**	read	stori	uniform	onlin	week	great	mooc	design	launch	use	READ
17	**1015**	**Know(ledge), good, (grade)**	know	thing	good	say	want	children	way	read	look	year	KNO
18	**125**	**Music, read, play**	music	read	stori	look	play	way	watch	word	know	bit	MUS
19	**42**	**Kinect, CLVfest**	kinect	http	applic	wall	day	read	view	english	clvfest	click	TECH

2013													
Topic Num	**Num Docs**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**	
0	**244**	**Knowledge, technology, (Memori(se)**	thing	way	idea	pupil	want	help	think	technolog	world	experi	KNO
1	**654**	**Class, word, resource**	class	use	word	look	resourc	number	idea	http	activ	ask	REC
2	**155**	**Language, learner, (ESOL, EFL) (Growth mindset)**	languag	learner	know	english	poem	start	read	idea	book	thing	LANG
3	**423**	**Ofsted, observation, inspector**	ofst	say	know	observ	thing	good	point	inspector	way	children	OFSTED
4	**46**	**Culture, pollito (Hay Festival)**	pollito	share	pio	cultur	list	hay	sort	use	radio	glu	CUL
5	**96**	**Music, certificate, edexcel, iGCSE**	music	certif	open	edexcel	workshop	english	igcs	group	support	quiz	MUS
6	**343**	**Music, science**	pupil	know	music	day	good	scienc	want	week	play	look	MUS
7	**250**	**Event, London, primary**	day	event	london	week	primari	children	great	comput	lectur	visit	EVENT
8	**317**	**App, ipad, technology**	app	ipad	technolog	video	share	tool	creat	use	onlin	digit	TECH
9	**1568**	**Question, (feedback, assess)**	pupil	write	question	class	way	use	group	idea	know	ask	ASS
10	**79**	**Language, book, corpora**	languag	book	differ	class	word	labour	look	corpora	way	set	LANG
11	**243**	**Word, language, text (Voynich manuscript)**	word	look	languag	produc	text	way	differ	use	articl	book	LANG
12	**70**	**Comput(er, ing), ICT, problem, sleep**	children	comput	level	problem	ict	program	sleep	challeng	algorithm	technolog	TECH
13	**83**	**Dyslexia**	dyslexia	countri	english	word	differ	class	children	group	mean	speak	DYS
14	**1712**	**Change, govern (policy, Gove)**	children	chang	good	year	govern	say	develop	educ	research	support	GOVE
15	**85**	**Design, team, client**	design	team	client	case	way	share	requir	world	idea	look	MISC
16	**404**	**Read, children, book**	read	children	book	love	want	thing	know	day	start	feel	READ
17	**91**	**Music, bulli(y, ing)**	music	boy	girl	bulli	elf	say	women	feel	know	way	MUS
18	**148**	**Memori(se), knowledge, cognit(ive) (Willingham, Hatti)**	memori	thing	knowledg	kid	research	cognit	thought	say	know	brain	MC
19	**47**	**Primary, Brentwood**	primari	brentwood	counti	local	hall	essex	ecc	council	sawyer	share	PRI

2014												
Topic Num	**Num Documents**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**
0	**1291**	**Feedback, question, mark**	question	feedback	mark	idea	write	class	plan	answer	use	ask
1	**131**	**Read, phonic, (decode, dyslexia)**	read	word	children	phonic	child	sound	reader	book	reading.	say
2	**280**	**Policy, govern, Gove**	polici	govern	children	report	gove	free	money	fund	pupil	improv
3	**203**	**Leader, colleague, leadership**	thing	leader	know	say	colleagu	staff	experi	good	leadership	feel
4	**1189**	**Knowledge, language, understand**	children	knowledg	understand	way	skill	languag	write	develop	differ	word
5	**1711**	**Thing, want, week**	thing	want	week	know	day	feel	class	children	read	help
6	**848**	**Ofsted, assess(ment), grade, progress**	ofst	assess	pupil	grade	good	progress	level	know	behaviour	say
7	**181**	**War, history**	play	know	war	outdoor	histori	world	book	day	littl	use
8	**38**	**Poem, poetry, Blake**	poem	trivium	pupil	nuclear	blake	idea	dialect	world	poetri	robinson
9	**102**	**College, sixth form, vocation(al), university (FE/HE)**	colleg	sixth	form	vocat	univers	qualif	young	london	progress	level
10	**623**	**Research, technology, ipad**	research	technolog	app	digit	ipad	onlin	use	way	access	develop
11	**264**	**Resource, open, create, share (Spanish)**	resourc	open	creat	share	activ	card	link	app	use	scotland
12	**131**	**Policy, academy, (admiss(ion))**	http	local	parent	author	free	polici	state	children	academi	evid
13	**237**	**Comput(er), image, code**	articl	comput	https	imag	use	look	differ	code	exampl	way
14	**97**	**Research, evidence, science, education, practice**	research	evid	scienc	educ	practic	theori	method	argument	claim	effect
15	**881**	**Love, know**	love	children	day	thing	know	look	littl	bit	go	start
16	**275**	**Read, book, story, character**	read	book	stori	text	charact	life	man	love	word	world
17	**193**	**Music, song, sing**	music	song	sing	twitter	play	present	share	les	tweet	listen
18	**296**	**Social, value, culture, society**	social	valu	cultur	educ	societi	world	young	knowledg	develop	way
19	**887**	**Staff, change, support**	staff	chang	support	want	children	work	good	year	day	job

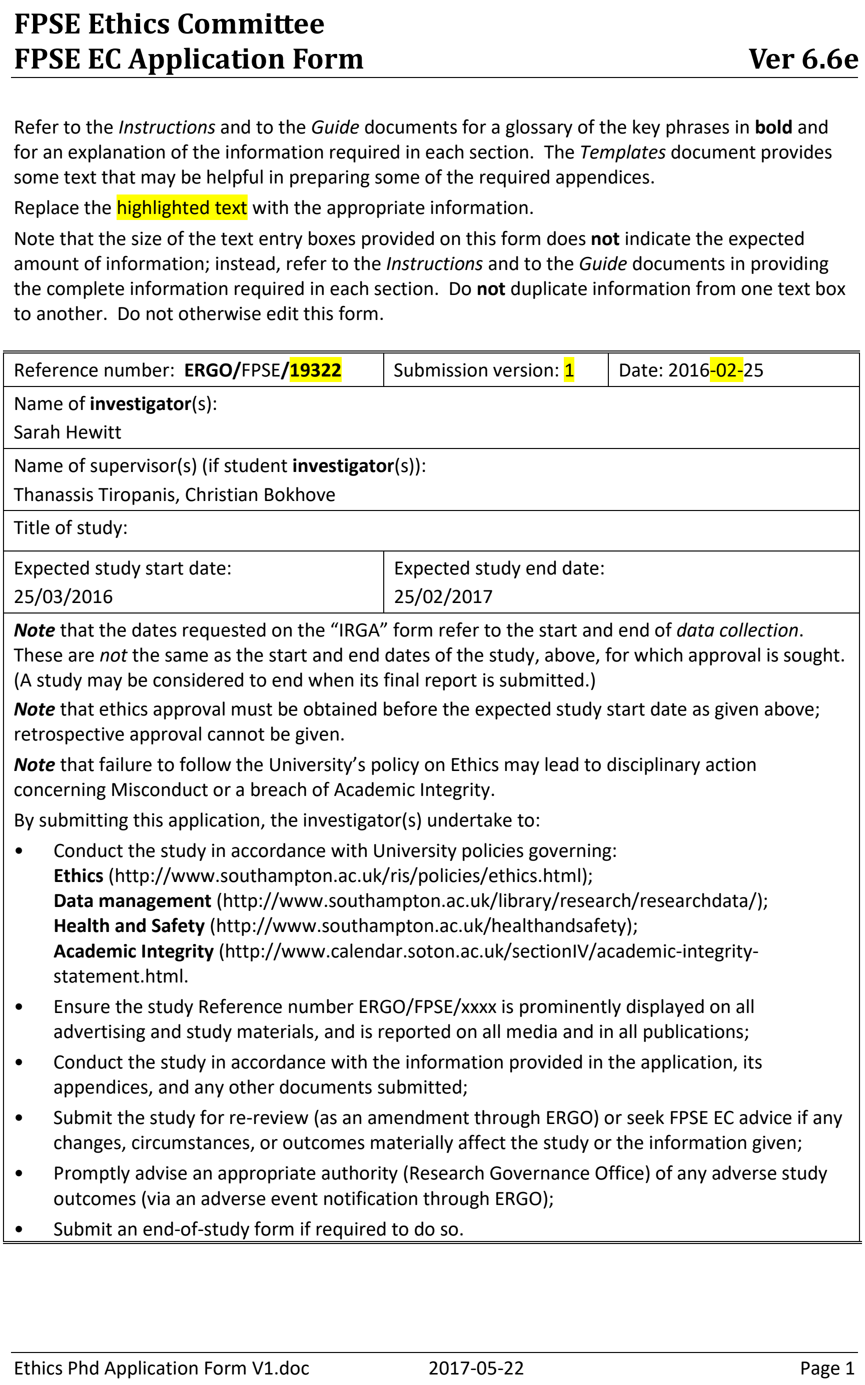
2015												
Topic Num	**Num Documents**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**
0	**422**	**Exam, mark, question (GCSE)**	way	exam	question	ask	mark	know	chang	differ	good	pupil
1	**258**	**Idea, education, help**	idea	educ	help	twitter	good	compani	look	way	social	know
2	**16**	**Game, drug**	dice	game	drug	draw	throw	play	egg	roll	scienc	poetri
3	**1150**	**Number, (math, resource)**	class	children	know	use	number	start	idea	question	set	way
4	**2338**	**Love, feel**	day	children	know	thing	love	feel	week	look	want	advertis
5	**139**	**Knowledge, memory, process, cognit(ive, ition)**	knowledg	memori	process	cognit	word	research	test	skill	effect	model
6	**110**	**Gender, language, study**	boy	english	languag	studi	read	british	gender	girl	research	issu
7	**83**	**OCR**	august	link	hexagon	octob	april	activ	ocr	joe	advertis	spedsc
8	**622**	**Knowledge, skill, research**	knowledg	skill	research	parti	social	govern	educ	polit	develop	young
9	**146**	**Country, culture, (Shakespeare)**	countri	cultur	know	play	world	say	human	want	man	coach
10	**48**	**Music, band, minecraft**	music	song	band	sing	minecraft	play	perform	instrument	music.	album
11	**162**	**Language, learner, listen (MFL)**	languag	learner	listen	activ	target	grammar	french	skill	ask	task
12	**569**	**Science, physic(s), math(s),**	scienc	physic	math	subject	project	research	sport	look	idea	digit
13	**75**	**College, sixth form, university (FE)**	colleg	sixth	univers	form	london	progress	market	select	cours	young
14	**73**	**Welsh**	mae	bod	les	wedi	hwn	hyn	iaith	wrth	yma	ond
15	**446**	**Free, research, academy**	children	parent	support	free	research	year	academi	fund	local	money
16	**2818**	**Develop, assess(ment), (progress)**	pupil	know	way	develop	assess	good	thing	children	help	want
17	**40**	**Primary, secondary, sport**	yes	pit	primari	youth	general	secondari	sport	isobel	danc	studio
18	**58**	**Week, function, assembly**	week	function	var	pupil	return	day	trip	number	assembl	word
19	**938**	**Read, write, book**	read	word	write	book	use	children	text	app	languag	creat

2016												
Topic Num	**Num Documents**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**
0	**285**	**Knowledge, music, art (creative)**	knowledg	music	know	pupil	say	idea	understand	art	way	thing
1	**78**	**Week, Catholic**	week	group	pupil	number	great	parent	cathol	welsh	word	rule
2	**168**	**Language, word, read, grammar**	languag	word	read	text	english	grammar	sentenc	listen	vocabulari	mean
3	**36**	**Music, game, sport**	music	game	sport	coach	play	way	nous	attach	des	les
4	**80**	**MAT, academy, board**	mat	trust	academi	way	board	use	look	say	thing	line
5	**66**	**Read, book, response, share**	read	pupil	children	book	respons	share	listen	challeng	charact	reader
6	**72**	**Character, faith, God**	charact	world	develop	live	valu	faith	virtu	life	god	young
7	**298**	**App, technology, computer**	app	digit	use	technolog	comput	creat	share	educ	idea	onlin
8	**419**	**Policy, govern(ment), education**	polici	govern	educ	read	support	chang	research	point	social	year
9	**59**	**Vote, politics, campaign, brexit**	vote	polit	campaign	immigr	referendum	brexit	war	european	tori	card
10	**60**	**Coach, pronunci(ation), phonem(e), iGCSE**	coach	video	quiz	pronunci	igcs	phonem	practic	data	collect	ensur
11	**1292**	**Feel, love**	thing	day	know	feel	want	say	look	love	children	come
12	**188**	**Assess, test, data, grade, progress, level**	assess	pupil	test	data	grade	progress	level	mark	result	expect
13	**907**	**Question, maths, respource**	question	math	use	resourc	write	mark	answer	help	class	plan
14	**118**	**Univers(al, e), health, mental**	univers	health	open	mental	music	london	children	cours	young	skill
15	**370**	**Research, inform(ation), effect(ive), practic€**	research	knowledg	inform	effect	understand	practic	way	learner	know	evid
16	**49**	**Sixth, form, college (FE)**	sixth	form	colleg	london	averag	borough	level	local	offer	fund
17	**118**	**Shakespeare, read, Michaela**	shakespear	book	read	world	idea	michaela	polit	english	macbeth	present
18	**1275**	**Children, staff (behaviour)**	children	pupil	way	know	help	staff	thing	support	child	work
19	**37**	**Welsh**	mae	bod	hwn	neu	wedi	wrth	roedd	disgyblion	dysgu	hyn

2017												
Topic Num	**Num Documents**		**Word 0**	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word 8**	**Word 9**
0	**290**	**Write, sentence, (vocabulary)**	word	write	text	read	use	languag	question	sentenc	task	exampl
1	**236**	**Play, (space, preschool)**	children	play	way	child	thing	idea	feel	differ	person	make
2	**37**	**Welsh**	mae	iaith	bod	book	hyn	mewn	gan	wrth	spanish	mae'r
3	**14**	**Corsica, Matisse**	corsica	corsican	matiss	paint	home	colour	stori	juli	masculin	idea
4	**115**	**World, polit(ics), women**	world	polit	human	read	say	social	book	power	women	thing
5	**46**	**London, Univers(al?), community**	london	univers	communiti	newvic	social	east	colleg	achiev	progress	organis
6	**297**	**Digit(al), music, tool, app**	digit	music	tool	languag	creat	free	comput	activ	app	develop
7	**87**	**University, college, (FE)**	univers	nation	sixth	colleg	form	london	progress	young	good	skill
8	**67**	**Book, read, literature, (OCR, Gatsby)**	book	read	stori	gatsbi	week	love	nick	literatur	write	chapter
9	**100**	**Revision, exam, question (EdExcel)**	exam	revis	question	topic	test	use	look	paper	studi	resourc
10	**37**	**Energy, store, family**	energi	store	famili	mother	physic	point	god	day	form	love
11	**196**	**Research, support, sport**	pupil	research	support	physic	provid	help	opportun	sport	local	fund
12	**136**	**Maths, problem, solve**	math	problem	mathemat	understand	solv	number	idea	way	equat	differ
13	**161**	**Fund, govern(ment), policy**	fund	govern	support	right	univers	nation	pay	local	cost	polici
14	**172**	**Resource, write, (website)**	week	resourc	write	say	look	number	day	read	pupil	primari
15	**357**	**Staff, plan, feedback, mark**	thing	staff	know	class	want	plan	work	mark	good	start
16	**129**	**Course, language (IATEFL 2017)**	april	question	help	activ	user	cours	languag	data	word	listen
17	**26**	**Staff, ISSU**	yes	staff	come	score	ident	week	issu	game	veo	year
18	**720**	**Knowledge, assess(ment), test, (progress)**	knowledg	assess	know	subject	pupil	way	children	differ	test	good
19	**446**	**Know, day, want, feel**	know	day	thing	want	children	feel	year	good	say	life

Appendix D

Ethics Approval



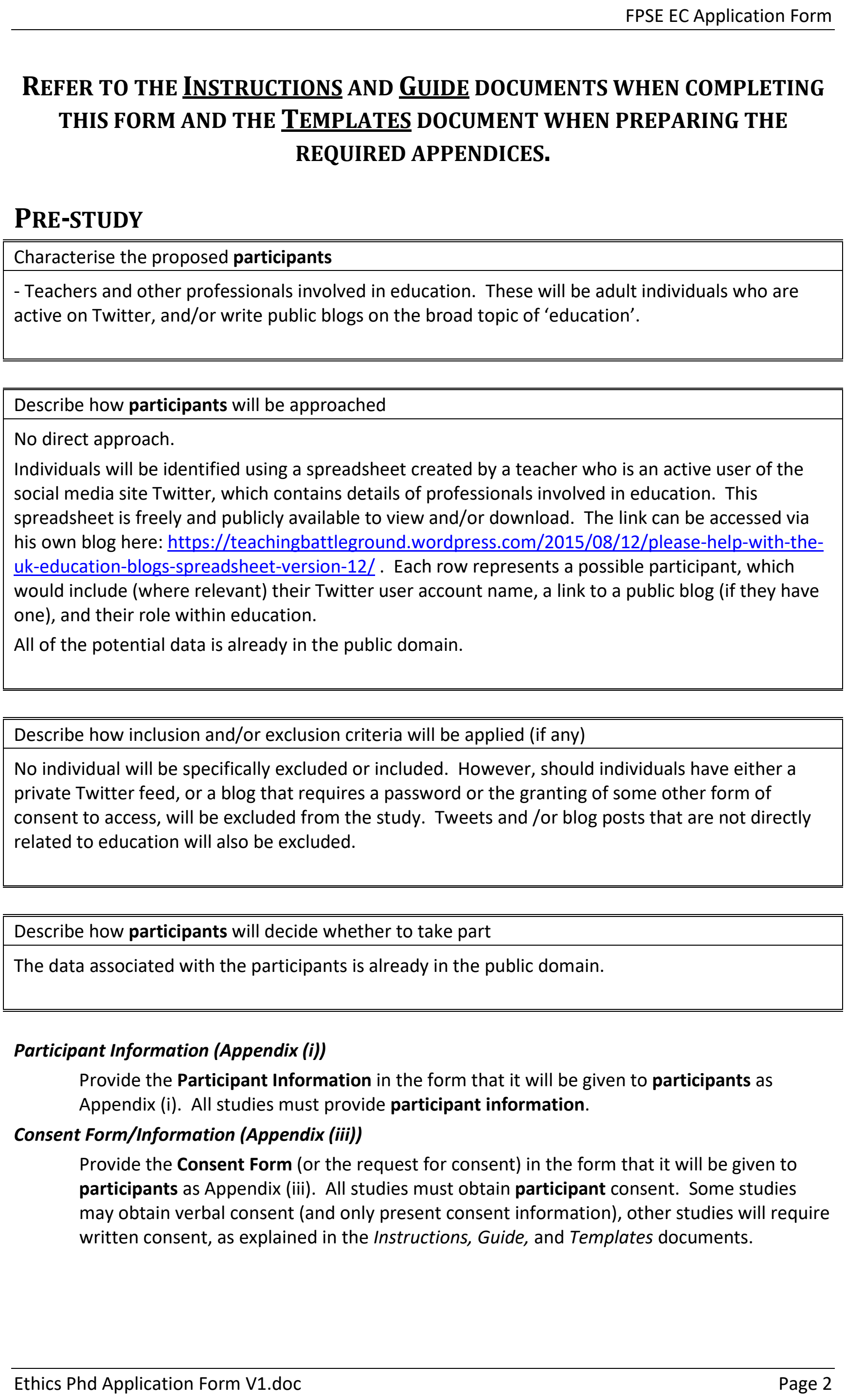
FPSE Ethics Committee
FPSE EC Application Form Ver 6.6e

Refer to the *Instructions* and to the *Guide* documents for a glossary of the key phrases in **bold** and for an explanation of the information required in each section. The *Templates* document provides some text that may be helpful in preparing some of the required appendices.

Replace the highlighted text with the appropriate information.

Note that the size of the text entry boxes provided on this form does **not** indicate the expected amount of information; instead, refer to the *Instructions* and to the *Guide* documents in providing the complete information required in each section. Do **not** duplicate information from one text box to another. Do not otherwise edit this form.

Reference number: **ERGO/**FPSE/**19322**	Submission version: 1	Date: 2016-02-25
Name of **investigator**(s): Sarah Hewitt		
Name of supervisor(s) (if student **investigator**(s)): Thanassis Tiropanis, Christian Bokhove		
Title of study:		
Expected study start date: 25/03/2016	Expected study end date: 25/02/2017	
Note that the dates requested on the "IRGA" form refer to the start and end of *data collection*. These are *not* the same as the start and end dates of the study, above, for which approval is sought. (A study may be considered to end when its final report is submitted.) ***Note*** that ethics approval must be obtained before the expected study start date as given above; retrospective approval cannot be given. ***Note*** that failure to follow the University's policy on Ethics may lead to disciplinary action concerning Misconduct or a breach of Academic Integrity. By submitting this application, the investigator(s) undertake to: • Conduct the study in accordance with University policies governing: **Ethics** (http://www.southampton.ac.uk/ris/policies/ethics.html); **Data management** (http://www.southampton.ac.uk/library/research/researchdata/); **Health and Safety** (http://www.southampton.ac.uk/healthandsafety); **Academic Integrity** (http://www.calendar.soton.ac.uk/sectionIV/academic-integrity-statement.html. • Ensure the study Reference number ERGO/FPSE/xxxx is prominently displayed on all advertising and study materials, and is reported on all media and in all publications; • Conduct the study in accordance with the information provided in the application, its appendices, and any other documents submitted; • Submit the study for re-review (as an amendment through ERGO) or seek FPSE EC advice if any changes, circumstances, or outcomes materially affect the study or the information given; • Promptly advise an appropriate authority (Research Governance Office) of any adverse study outcomes (via an adverse event notification through ERGO); • Submit an end-of-study form if required to do so.		



REFER TO THE <u>INSTRUCTIONS</u> AND <u>GUIDE</u> DOCUMENTS WHEN COMPLETING THIS FORM AND THE <u>TEMPLATES</u> DOCUMENT WHEN PREPARING THE REQUIRED APPENDICES.

PRE-STUDY

Characterise the proposed **participants**

- Teachers and other professionals involved in education. These will be adult individuals who are active on Twitter, and/or write public blogs on the broad topic of 'education'.

Describe how **participants** will be approached

No direct approach.

Individuals will be identified using a spreadsheet created by a teacher who is an active user of the social media site Twitter, which contains details of professionals involved in education. This spreadsheet is freely and publicly available to view and/or download. The link can be accessed via his own blog here: https://teachingbattleground.wordpress.com/2015/08/12/please-help-with-the-uk-education-blogs-spreadsheet-version-12/ . Each row represents a possible participant, which would include (where relevant) their Twitter user account name, a link to a public blog (if they have one), and their role within education.

All of the potential data is already in the public domain.

Describe how inclusion and/or exclusion criteria will be applied (if any)

No individual will be specifically excluded or included. However, should individuals have either a private Twitter feed, or a blog that requires a password or the granting of some other form of consent to access, will be excluded from the study. Tweets and /or blog posts that are not directly related to education will also be excluded.

Describe how **participants** will decide whether to take part

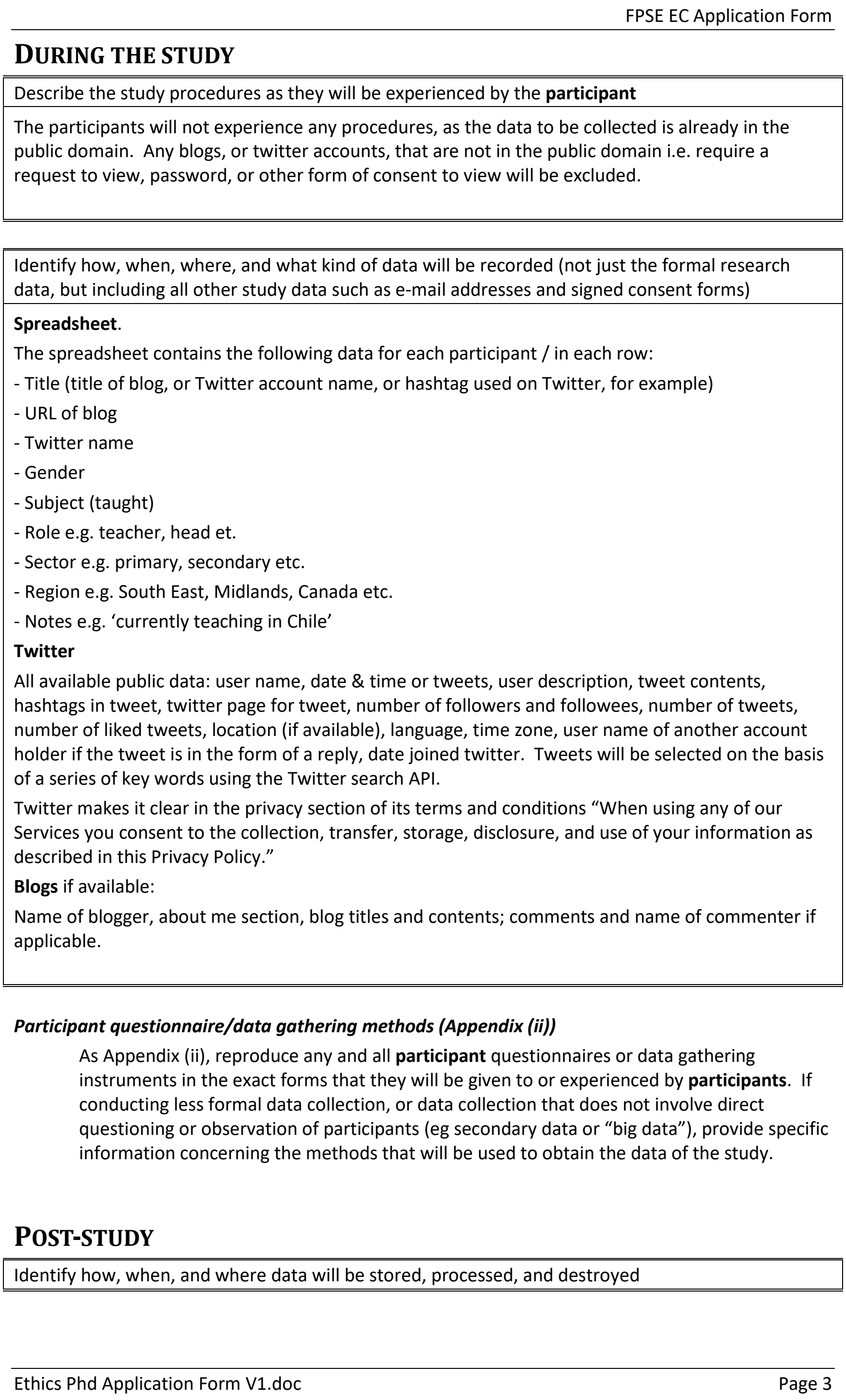
The data associated with the participants is already in the public domain.

Participant Information (Appendix (i))

Provide the **Participant Information** in the form that it will be given to **participants** as Appendix (i). All studies must provide **participant information**.

Consent Form/Information (Appendix (iii))

Provide the **Consent Form** (or the request for consent) in the form that it will be given to **participants** as Appendix (iii). All studies must obtain **participant** consent. Some studies may obtain verbal consent (and only present consent information), other studies will require written consent, as explained in the *Instructions, Guide,* and *Templates* documents.



DURING THE STUDY

Describe the study procedures as they will be experienced by the **participant**

The participants will not experience any procedures, as the data to be collected is already in the public domain. Any blogs, or twitter accounts, that are not in the public domain i.e. require a request to view, password, or other form of consent to view will be excluded.

Identify how, when, where, and what kind of data will be recorded (not just the formal research data, but including all other study data such as e-mail addresses and signed consent forms)

Spreadsheet.

The spreadsheet contains the following data for each participant / in each row:

- Title (title of blog, or Twitter account name, or hashtag used on Twitter, for example)
- URL of blog
- Twitter name
- Gender
- Subject (taught)
- Role e.g. teacher, head et.
- Sector e.g. primary, secondary etc.
- Region e.g. South East, Midlands, Canada etc.
- Notes e.g. ‘currently teaching in Chile’

Twitter

All available public data: user name, date & time or tweets, user description, tweet contents, hashtags in tweet, twitter page for tweet, number of followers and followees, number of tweets, number of liked tweets, location (if available), language, time zone, user name of another account holder if the tweet is in the form of a reply, date joined twitter. Tweets will be selected on the basis of a series of key words using the Twitter search API.

Twitter makes it clear in the privacy section of its terms and conditions “When using any of our Services you consent to the collection, transfer, storage, disclosure, and use of your information as described in this Privacy Policy.”

Blogs if available:

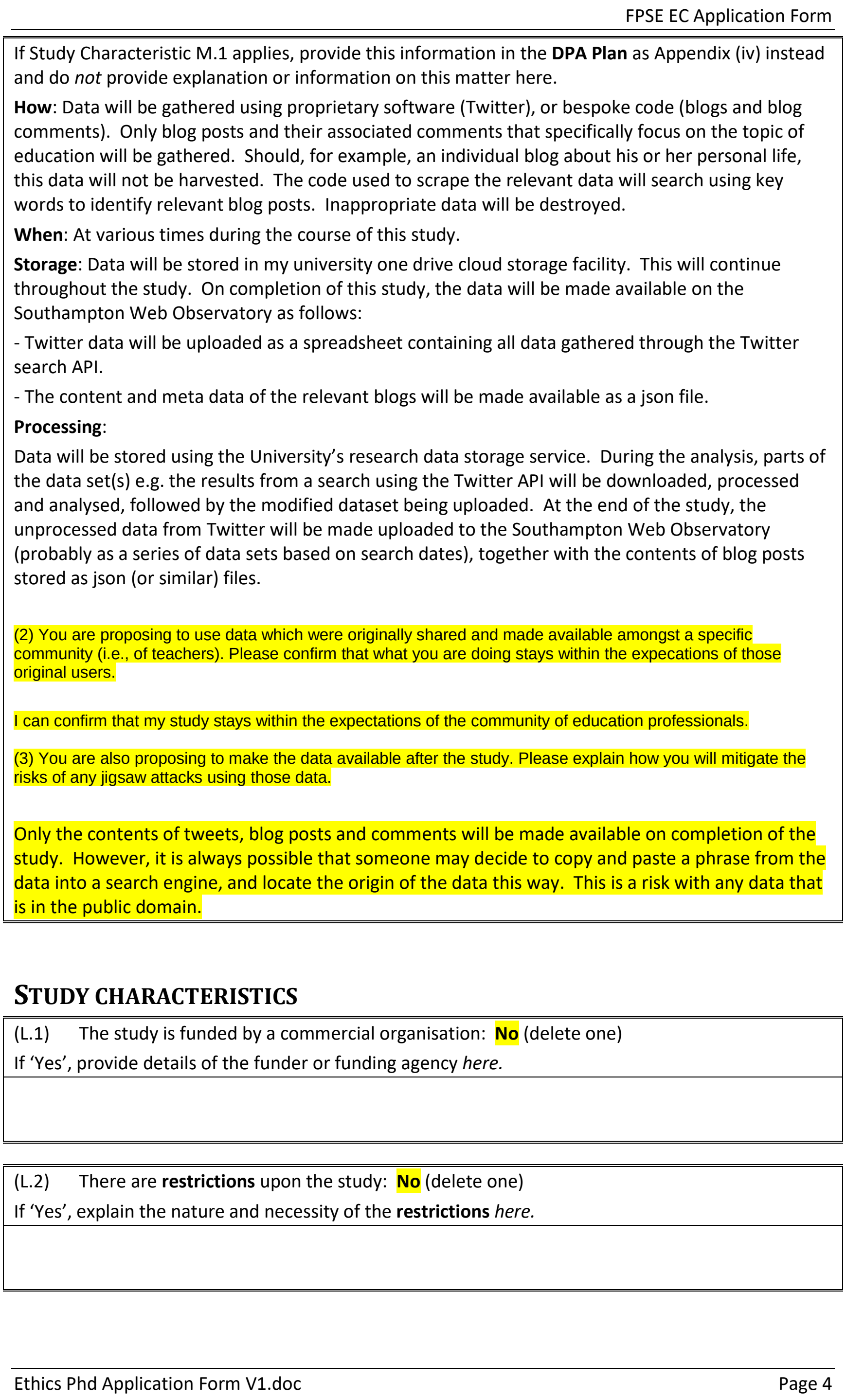
Name of blogger, about me section, blog titles and contents; comments and name of commenter if applicable.

Participant questionnaire/data gathering methods (Appendix (ii))

As Appendix (ii), reproduce any and all **participant** questionnaires or data gathering instruments in the exact forms that they will be given to or experienced by **participants**. If conducting less formal data collection, or data collection that does not involve direct questioning or observation of participants (eg secondary data or “big data”), provide specific information concerning the methods that will be used to obtain the data of the study.

POST-STUDY

Identify how, when, and where data will be stored, processed, and destroyed



If Study Characteristic M.1 applies, provide this information in the **DPA Plan** as Appendix (iv) instead and do *not* provide explanation or information on this matter here.

How: Data will be gathered using proprietary software (Twitter), or bespoke code (blogs and blog comments). Only blog posts and their associated comments that specifically focus on the topic of education will be gathered. Should, for example, an individual blog about his or her personal life, this data will not be harvested. The code used to scrape the relevant data will search using key words to identify relevant blog posts. Inappropriate data will be destroyed.

When: At various times during the course of this study.

Storage: Data will be stored in my university one drive cloud storage facility. This will continue throughout the study. On completion of this study, the data will be made available on the Southampton Web Observatory as follows:

- Twitter data will be uploaded as a spreadsheet containing all data gathered through the Twitter search API.

- The content and meta data of the relevant blogs will be made available as a json file.

Processing:

Data will be stored using the University's research data storage service. During the analysis, parts of the data set(s) e.g. the results from a search using the Twitter API will be downloaded, processed and analysed, followed by the modified dataset being uploaded. At the end of the study, the unprocessed data from Twitter will be made uploaded to the Southampton Web Observatory (probably as a series of data sets based on search dates), together with the contents of blog posts stored as json (or similar) files.

(2) You are proposing to use data which were originally shared and made available amongst a specific community (i.e., of teachers). Please confirm that what you are doing stays within the expecations of those original users.

I can confirm that my study stays within the expectations of the community of education professionals.

(3) You are also proposing to make the data available after the study. Please explain how you will mitigate the risks of any jigsaw attacks using those data.

Only the contents of tweets, blog posts and comments will be made available on completion of the study. However, it is always possible that someone may decide to copy and paste a phrase from the data into a search engine, and locate the origin of the data this way. This is a risk with any data that is in the public domain.

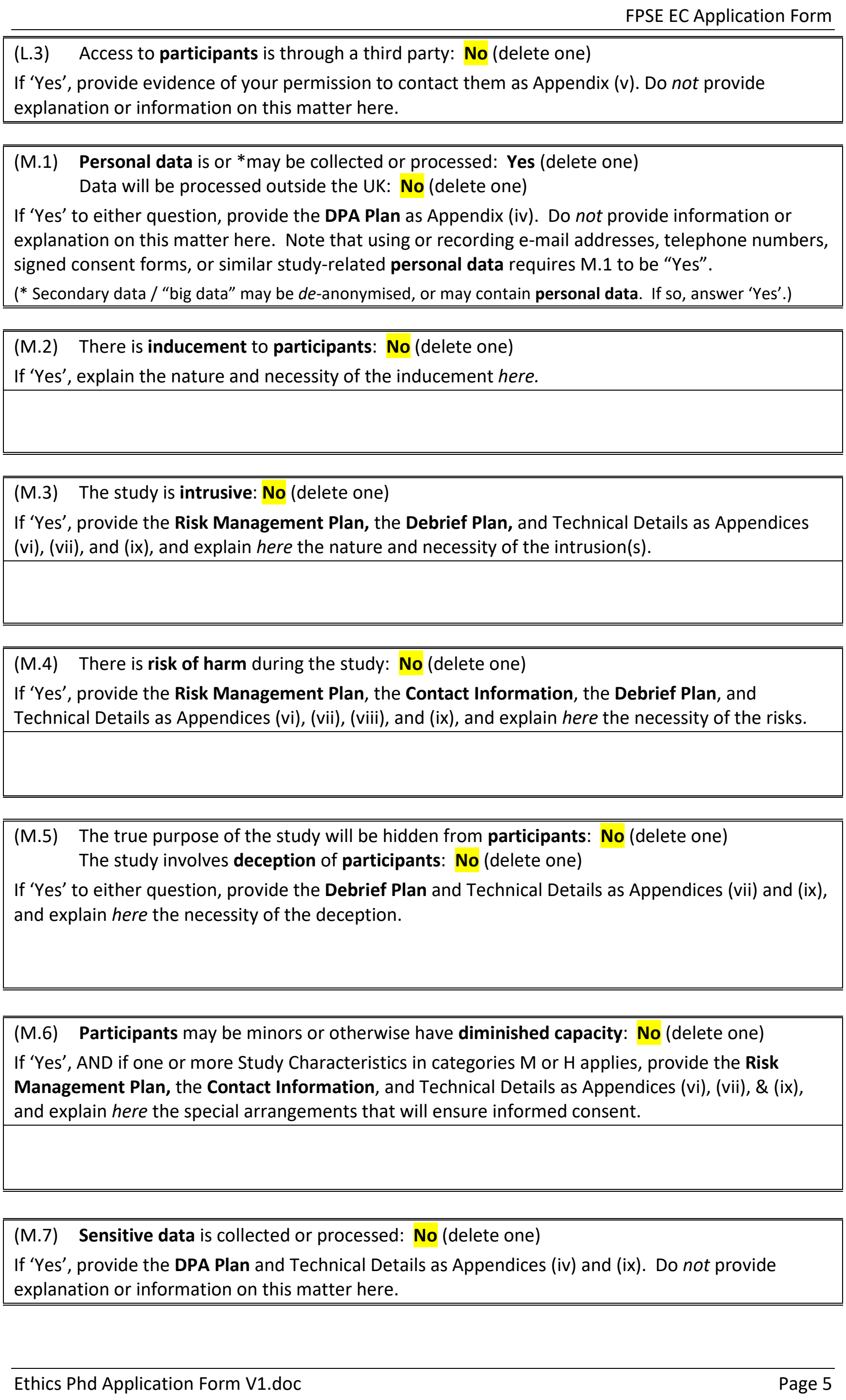
STUDY CHARACTERISTICS

(L.1) The study is funded by a commercial organisation: **No** (delete one)

If 'Yes', provide details of the funder or funding agency *here.*

(L.2) There are **restrictions** upon the study: **No** (delete one)

If 'Yes', explain the nature and necessity of the **restrictions** *here.*



(L.3) Access to **participants** is through a third party: **No** (delete one)

If 'Yes', provide evidence of your permission to contact them as Appendix (v). Do *not* provide explanation or information on this matter here.

(M.1) **Personal data** is or *may be collected or processed: **Yes** (delete one)
Data will be processed outside the UK: **No** (delete one)

If 'Yes' to either question, provide the **DPA Plan** as Appendix (iv). Do *not* provide information or explanation on this matter here. Note that using or recording e-mail addresses, telephone numbers, signed consent forms, or similar study-related **personal data** requires M.1 to be "Yes".

(* Secondary data / "big data" may be *de*-anonymised, or may contain **personal data**. If so, answer 'Yes'.)

(M.2) There is **inducement** to **participants**: **No** (delete one)

If 'Yes', explain the nature and necessity of the inducement *here.*

(M.3) The study is **intrusive**: **No** (delete one)

If 'Yes', provide the **Risk Management Plan,** the **Debrief Plan,** and Technical Details as Appendices (vi), (vii), and (ix), and explain *here* the nature and necessity of the intrusion(s).

(M.4) There is **risk of harm** during the study: **No** (delete one)

If 'Yes', provide the **Risk Management Plan**, the **Contact Information**, the **Debrief Plan**, and Technical Details as Appendices (vi), (vii), (viii), and (ix), and explain *here* the necessity of the risks.

(M.5) The true purpose of the study will be hidden from **participants**: **No** (delete one)
The study involves **deception** of **participants**: **No** (delete one)

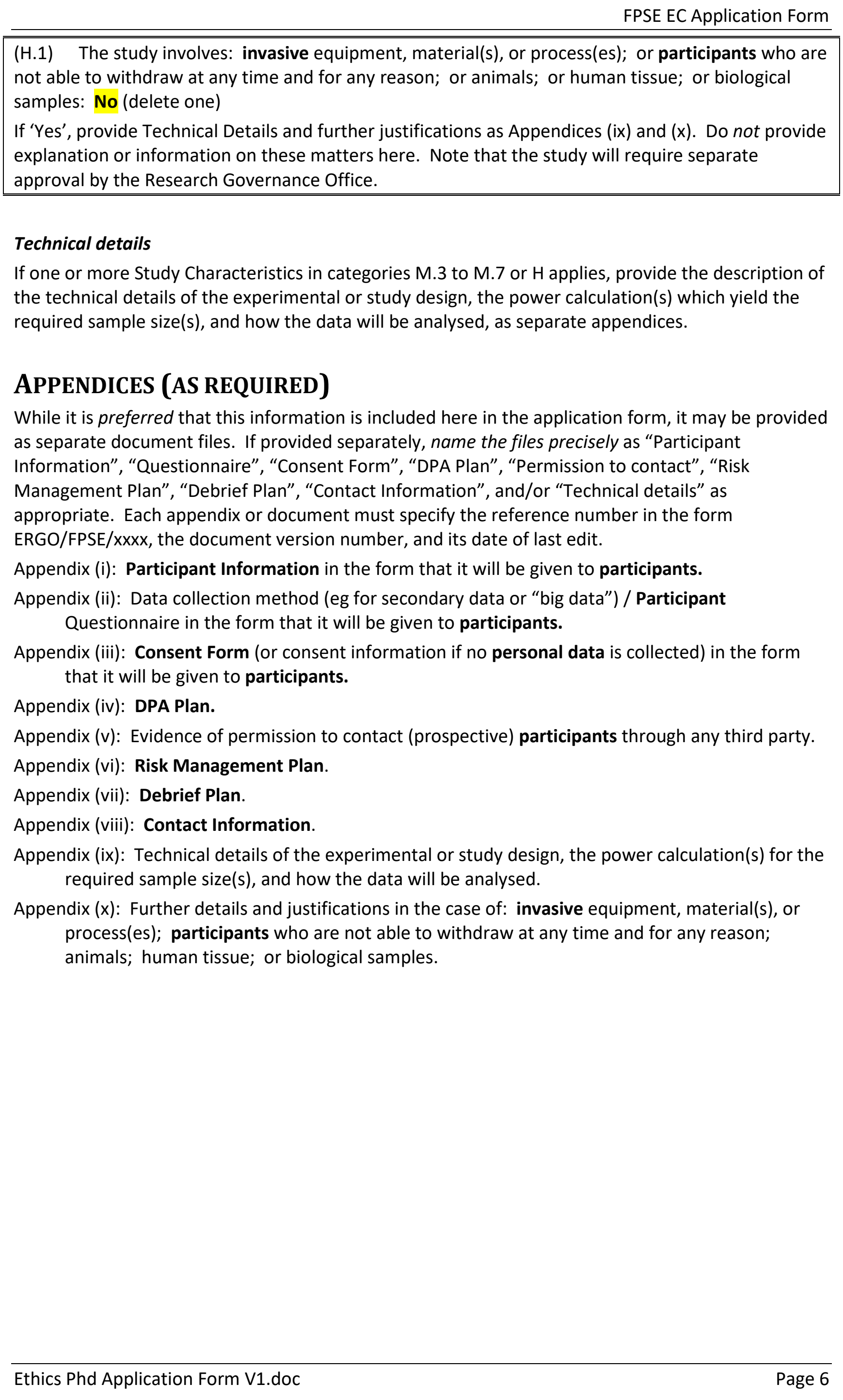
If 'Yes' to either question, provide the **Debrief Plan** and Technical Details as Appendices (vii) and (ix), and explain *here* the necessity of the deception.

(M.6) **Participants** may be minors or otherwise have **diminished capacity**: **No** (delete one)

If 'Yes', AND if one or more Study Characteristics in categories M or H applies, provide the **Risk Management Plan,** the **Contact Information**, and Technical Details as Appendices (vi), (vii), & (ix), and explain *here* the special arrangements that will ensure informed consent.

(M.7) **Sensitive data** is collected or processed: **No** (delete one)

If 'Yes', provide the **DPA Plan** and Technical Details as Appendices (iv) and (ix). Do *not* provide explanation or information on this matter here.



(H.1) The study involves: **invasive** equipment, material(s), or process(es); or **participants** who are not able to withdraw at any time and for any reason; or animals; or human tissue; or biological samples: **No** (delete one) If 'Yes', provide Technical Details and further justifications as Appendices (ix) and (x). Do *not* provide explanation or information on these matters here. Note that the study will require separate approval by the Research Governance Office.

Technical details

If one or more Study Characteristics in categories M.3 to M.7 or H applies, provide the description of the technical details of the experimental or study design, the power calculation(s) which yield the required sample size(s), and how the data will be analysed, as separate appendices.

APPENDICES (AS REQUIRED)

While it is *preferred* that this information is included here in the application form, it may be provided as separate document files. If provided separately, *name the files precisely* as "Participant Information", "Questionnaire", "Consent Form", "DPA Plan", "Permission to contact", "Risk Management Plan", "Debrief Plan", "Contact Information", and/or "Technical details" as appropriate. Each appendix or document must specify the reference number in the form ERGO/FPSE/xxxx, the document version number, and its date of last edit.

Appendix (i): **Participant Information** in the form that it will be given to **participants.**

Appendix (ii): Data collection method (eg for secondary data or "big data") / **Participant** Questionnaire in the form that it will be given to **participants.**

Appendix (iii): **Consent Form** (or consent information if no **personal data** is collected) in the form that it will be given to **participants.**

Appendix (iv): **DPA Plan.**

Appendix (v): Evidence of permission to contact (prospective) **participants** through any third party.

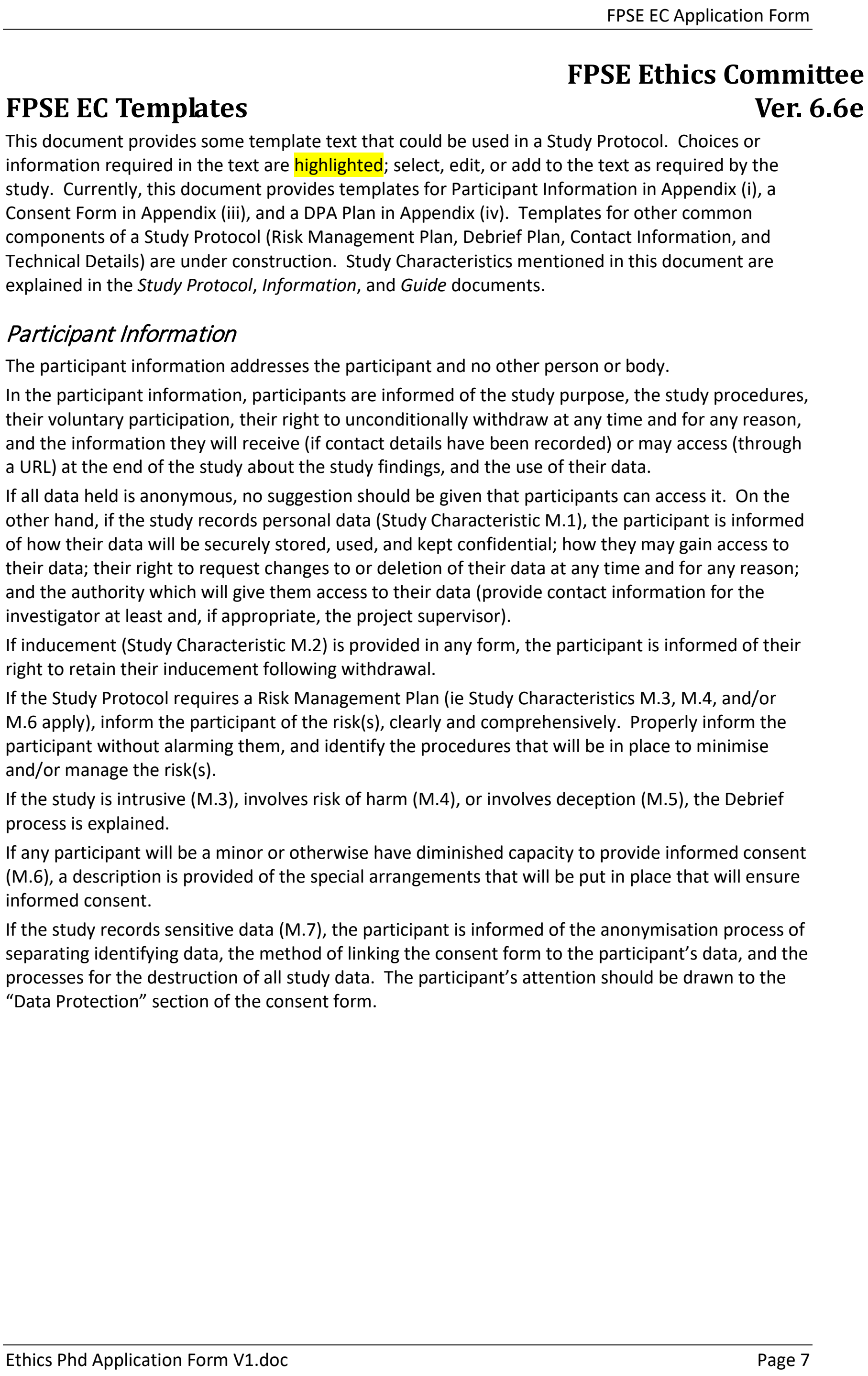
Appendix (vi): **Risk Management Plan**.

Appendix (vii): **Debrief Plan**.

Appendix (viii): **Contact Information**.

Appendix (ix): Technical details of the experimental or study design, the power calculation(s) for the required sample size(s), and how the data will be analysed.

Appendix (x): Further details and justifications in the case of: **invasive** equipment, material(s), or process(es); **participants** who are not able to withdraw at any time and for any reason; animals; human tissue; or biological samples.



FPSE Ethics Committee

FPSE EC Templates Ver. 6.6e

This document provides some template text that could be used in a Study Protocol. Choices or information required in the text are highlighted; select, edit, or add to the text as required by the study. Currently, this document provides templates for Participant Information in Appendix (i), a Consent Form in Appendix (iii), and a DPA Plan in Appendix (iv). Templates for other common components of a Study Protocol (Risk Management Plan, Debrief Plan, Contact Information, and Technical Details) are under construction. Study Characteristics mentioned in this document are explained in the *Study Protocol*, *Information*, and *Guide* documents.

Participant Information

The participant information addresses the participant and no other person or body.

In the participant information, participants are informed of the study purpose, the study procedures, their voluntary participation, their right to unconditionally withdraw at any time and for any reason, and the information they will receive (if contact details have been recorded) or may access (through a URL) at the end of the study about the study findings, and the use of their data.

If all data held is anonymous, no suggestion should be given that participants can access it. On the other hand, if the study records personal data (Study Characteristic M.1), the participant is informed of how their data will be securely stored, used, and kept confidential; how they may gain access to their data; their right to request changes to or deletion of their data at any time and for any reason; and the authority which will give them access to their data (provide contact information for the investigator at least and, if appropriate, the project supervisor).

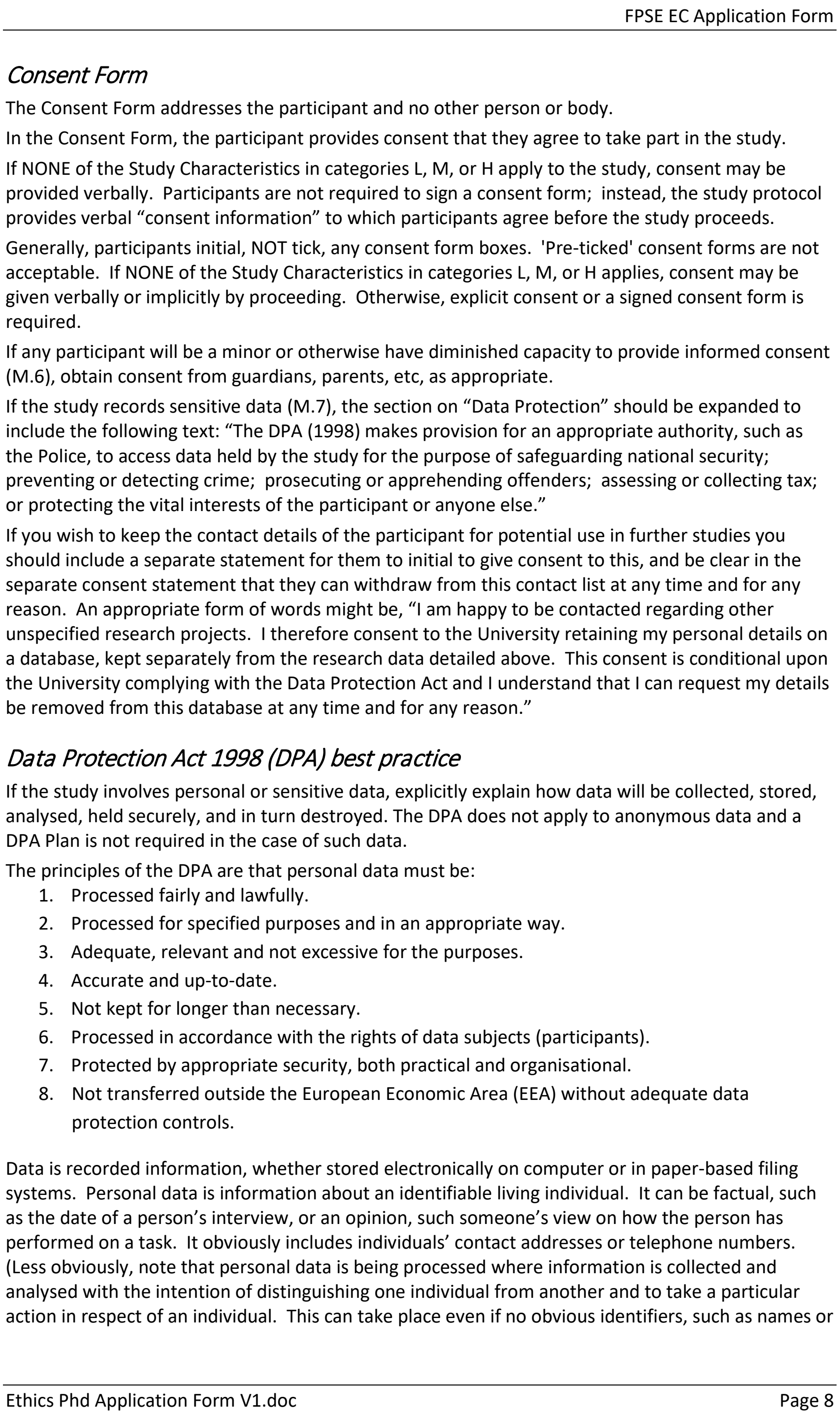
If inducement (Study Characteristic M.2) is provided in any form, the participant is informed of their right to retain their inducement following withdrawal.

If the Study Protocol requires a Risk Management Plan (ie Study Characteristics M.3, M.4, and/or M.6 apply), inform the participant of the risk(s), clearly and comprehensively. Properly inform the participant without alarming them, and identify the procedures that will be in place to minimise and/or manage the risk(s).

If the study is intrusive (M.3), involves risk of harm (M.4), or involves deception (M.5), the Debrief process is explained.

If any participant will be a minor or otherwise have diminished capacity to provide informed consent (M.6), a description is provided of the special arrangements that will be put in place that will ensure informed consent.

If the study records sensitive data (M.7), the participant is informed of the anonymisation process of separating identifying data, the method of linking the consent form to the participant's data, and the processes for the destruction of all study data. The participant's attention should be drawn to the "Data Protection" section of the consent form.



Consent Form

The Consent Form addresses the participant and no other person or body.

In the Consent Form, the participant provides consent that they agree to take part in the study.

If NONE of the Study Characteristics in categories L, M, or H apply to the study, consent may be provided verbally. Participants are not required to sign a consent form; instead, the study protocol provides verbal "consent information" to which participants agree before the study proceeds.

Generally, participants initial, NOT tick, any consent form boxes. 'Pre-ticked' consent forms are not acceptable. If NONE of the Study Characteristics in categories L, M, or H applies, consent may be given verbally or implicitly by proceeding. Otherwise, explicit consent or a signed consent form is required.

If any participant will be a minor or otherwise have diminished capacity to provide informed consent (M.6), obtain consent from guardians, parents, etc, as appropriate.

If the study records sensitive data (M.7), the section on "Data Protection" should be expanded to include the following text: "The DPA (1998) makes provision for an appropriate authority, such as the Police, to access data held by the study for the purpose of safeguarding national security; preventing or detecting crime; prosecuting or apprehending offenders; assessing or collecting tax; or protecting the vital interests of the participant or anyone else."

If you wish to keep the contact details of the participant for potential use in further studies you should include a separate statement for them to initial to give consent to this, and be clear in the separate consent statement that they can withdraw from this contact list at any time and for any reason. An appropriate form of words might be, "I am happy to be contacted regarding other unspecified research projects. I therefore consent to the University retaining my personal details on a database, kept separately from the research data detailed above. This consent is conditional upon the University complying with the Data Protection Act and I understand that I can request my details be removed from this database at any time and for any reason."

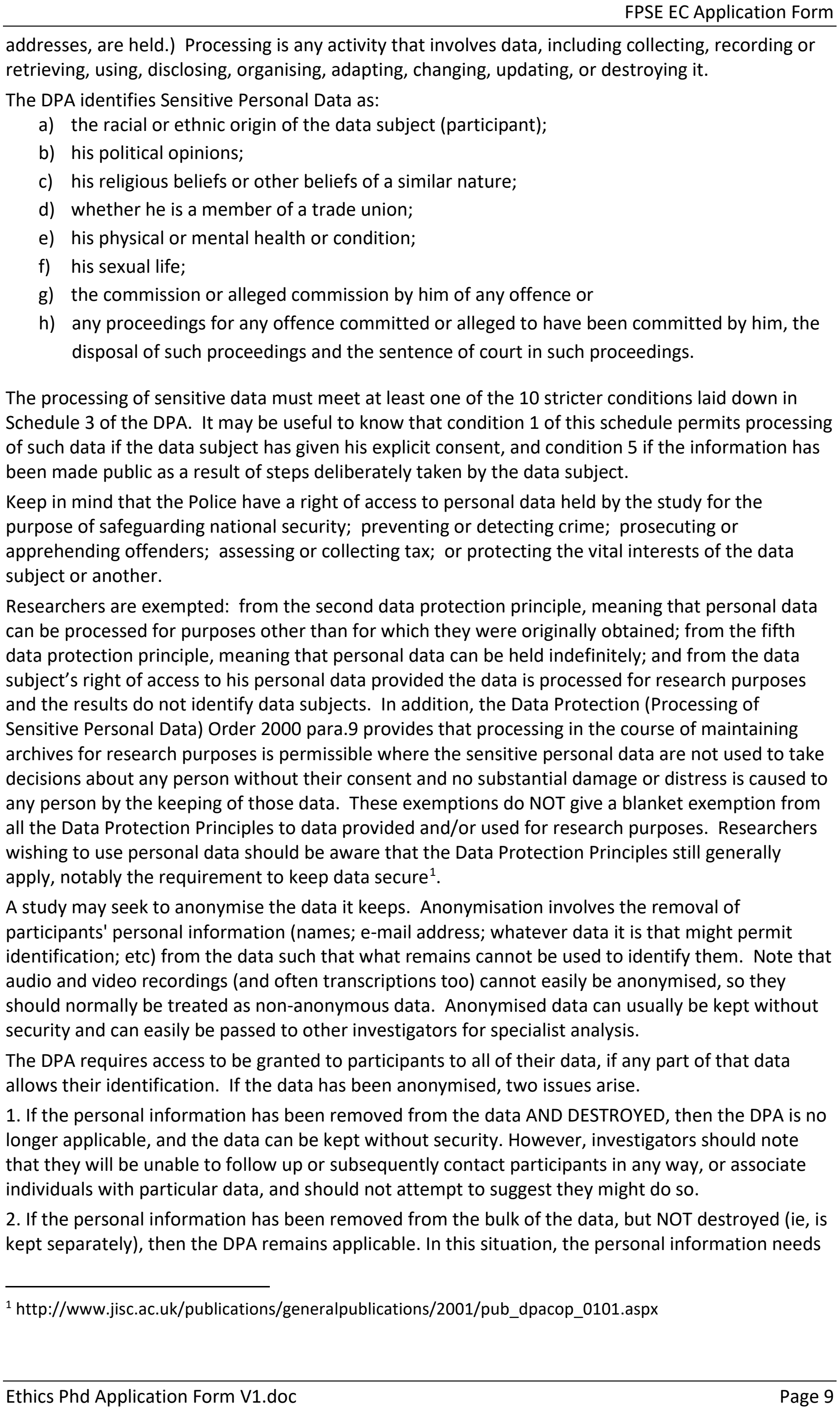
Data Protection Act 1998 (DPA) best practice

If the study involves personal or sensitive data, explicitly explain how data will be collected, stored, analysed, held securely, and in turn destroyed. The DPA does not apply to anonymous data and a DPA Plan is not required in the case of such data.

The principles of the DPA are that personal data must be:

1. Processed fairly and lawfully.
2. Processed for specified purposes and in an appropriate way.
3. Adequate, relevant and not excessive for the purposes.
4. Accurate and up-to-date.
5. Not kept for longer than necessary.
6. Processed in accordance with the rights of data subjects (participants).
7. Protected by appropriate security, both practical and organisational.
8. Not transferred outside the European Economic Area (EEA) without adequate data protection controls.

Data is recorded information, whether stored electronically on computer or in paper-based filing systems. Personal data is information about an identifiable living individual. It can be factual, such as the date of a person's interview, or an opinion, such someone's view on how the person has performed on a task. It obviously includes individuals' contact addresses or telephone numbers. (Less obviously, note that personal data is being processed where information is collected and analysed with the intention of distinguishing one individual from another and to take a particular action in respect of an individual. This can take place even if no obvious identifiers, such as names or



addresses, are held.) Processing is any activity that involves data, including collecting, recording or retrieving, using, disclosing, organising, adapting, changing, updating, or destroying it.

The DPA identifies Sensitive Personal Data as:

a) the racial or ethnic origin of the data subject (participant);
b) his political opinions;
c) his religious beliefs or other beliefs of a similar nature;
d) whether he is a member of a trade union;
e) his physical or mental health or condition;
f) his sexual life;
g) the commission or alleged commission by him of any offence or
h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings and the sentence of court in such proceedings.

The processing of sensitive data must meet at least one of the 10 stricter conditions laid down in Schedule 3 of the DPA. It may be useful to know that condition 1 of this schedule permits processing of such data if the data subject has given his explicit consent, and condition 5 if the information has been made public as a result of steps deliberately taken by the data subject.

Keep in mind that the Police have a right of access to personal data held by the study for the purpose of safeguarding national security; preventing or detecting crime; prosecuting or apprehending offenders; assessing or collecting tax; or protecting the vital interests of the data subject or another.

Researchers are exempted: from the second data protection principle, meaning that personal data can be processed for purposes other than for which they were originally obtained; from the fifth data protection principle, meaning that personal data can be held indefinitely; and from the data subject's right of access to his personal data provided the data is processed for research purposes and the results do not identify data subjects. In addition, the Data Protection (Processing of Sensitive Personal Data) Order 2000 para.9 provides that processing in the course of maintaining archives for research purposes is permissible where the sensitive personal data are not used to take decisions about any person without their consent and no substantial damage or distress is caused to any person by the keeping of those data. These exemptions do NOT give a blanket exemption from all the Data Protection Principles to data provided and/or used for research purposes. Researchers wishing to use personal data should be aware that the Data Protection Principles still generally apply, notably the requirement to keep data secure[1].

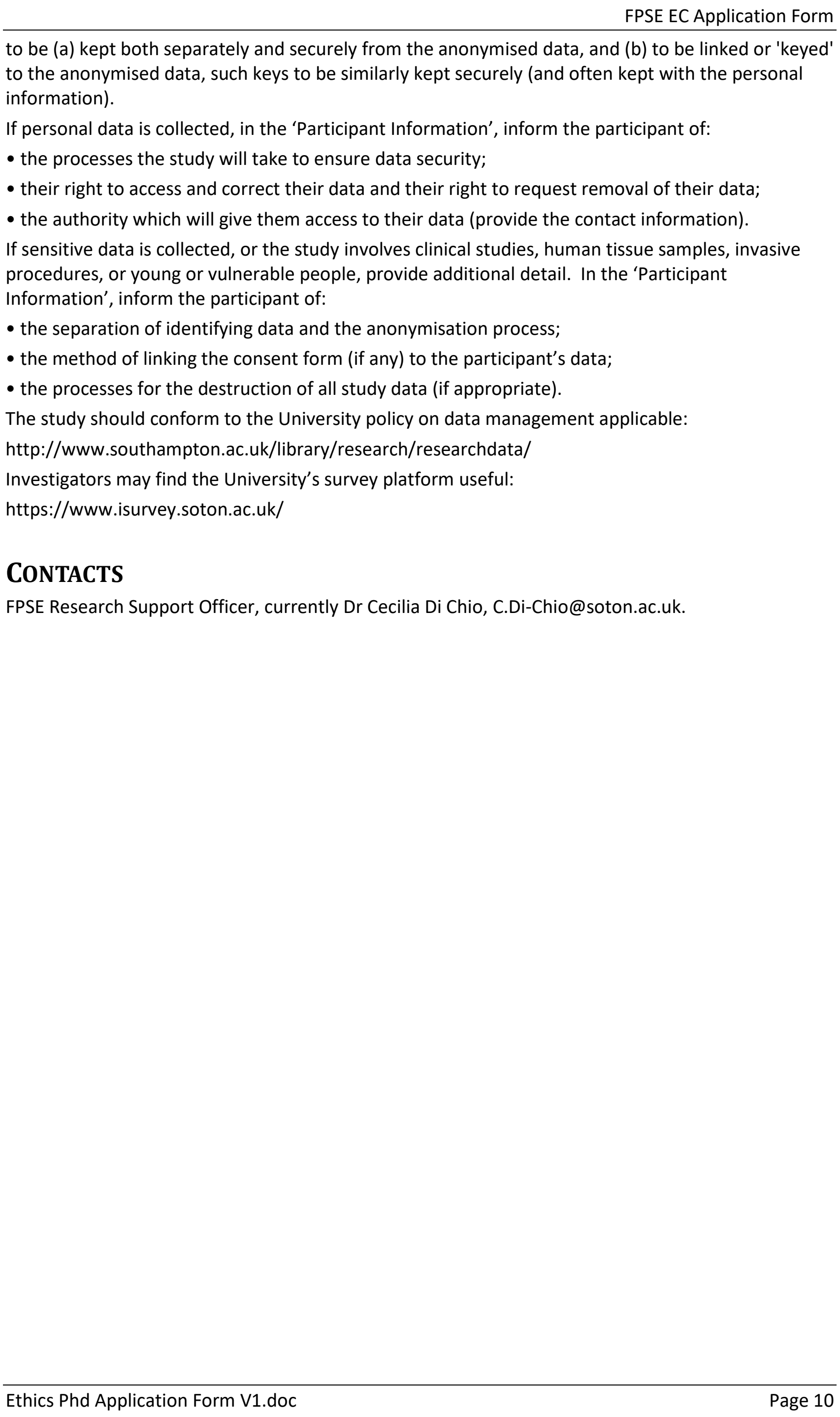
A study may seek to anonymise the data it keeps. Anonymisation involves the removal of participants' personal information (names; e-mail address; whatever data it is that might permit identification; etc) from the data such that what remains cannot be used to identify them. Note that audio and video recordings (and often transcriptions too) cannot easily be anonymised, so they should normally be treated as non-anonymous data. Anonymised data can usually be kept without security and can easily be passed to other investigators for specialist analysis.

The DPA requires access to be granted to participants to all of their data, if any part of that data allows their identification. If the data has been anonymised, two issues arise.

1. If the personal information has been removed from the data AND DESTROYED, then the DPA is no longer applicable, and the data can be kept without security. However, investigators should note that they will be unable to follow up or subsequently contact participants in any way, or associate individuals with particular data, and should not attempt to suggest they might do so.

2. If the personal information has been removed from the bulk of the data, but NOT destroyed (ie, is kept separately), then the DPA remains applicable. In this situation, the personal information needs

[1] http://www.jisc.ac.uk/publications/generalpublications/2001/pub_dpacop_0101.aspx



FPSE EC Application Form

to be (a) kept both separately and securely from the anonymised data, and (b) to be linked or 'keyed' to the anonymised data, such keys to be similarly kept securely (and often kept with the personal information).

If personal data is collected, in the ‘Participant Information’, inform the participant of:

- the processes the study will take to ensure data security;
- their right to access and correct their data and their right to request removal of their data;
- the authority which will give them access to their data (provide the contact information).

If sensitive data is collected, or the study involves clinical studies, human tissue samples, invasive procedures, or young or vulnerable people, provide additional detail. In the ‘Participant Information’, inform the participant of:

- the separation of identifying data and the anonymisation process;
- the method of linking the consent form (if any) to the participant’s data;
- the processes for the destruction of all study data (if appropriate).

The study should conform to the University policy on data management applicable:

http://www.southampton.ac.uk/library/research/researchdata/

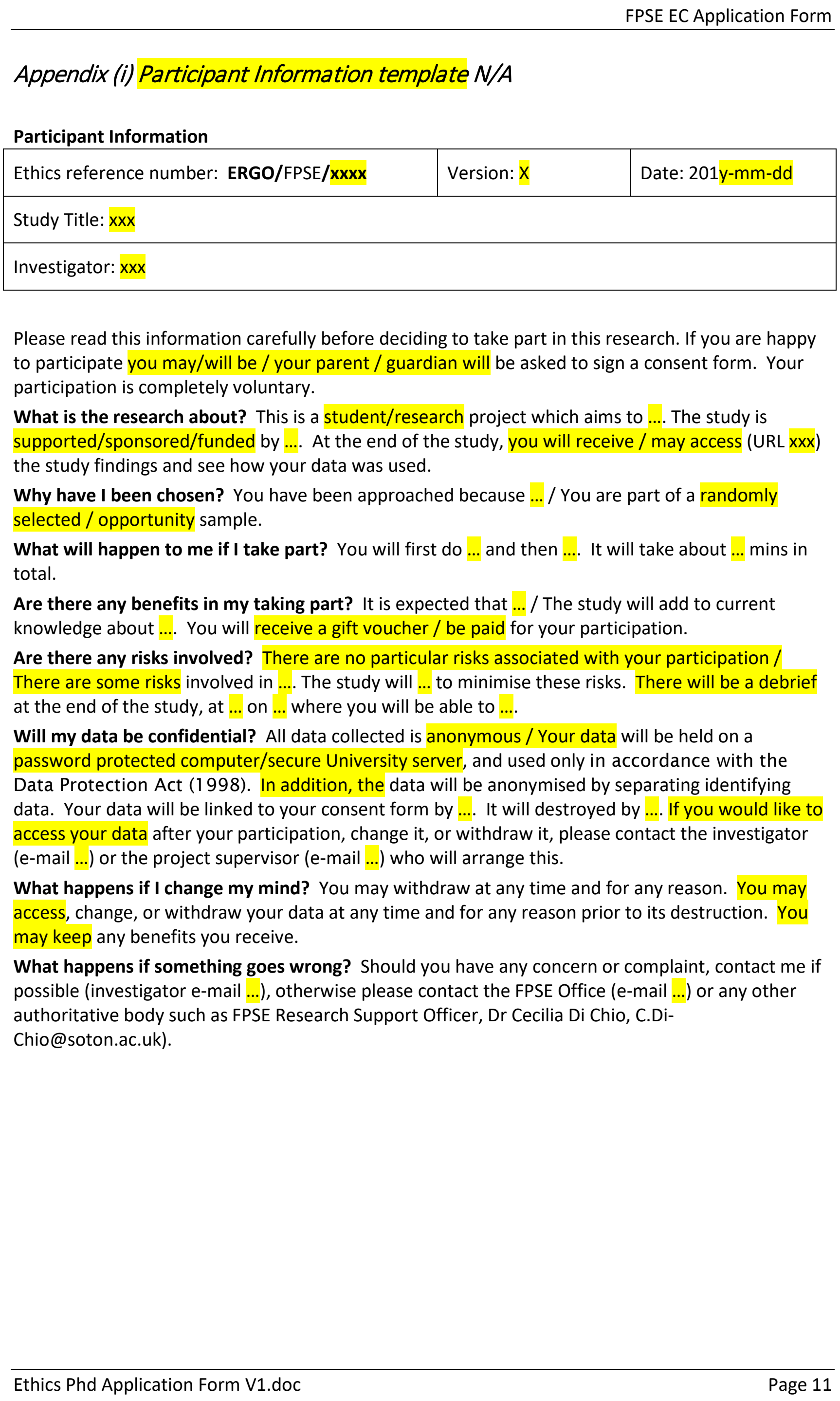
Investigators may find the University’s survey platform useful:

https://www.isurvey.soton.ac.uk/

CONTACTS

FPSE Research Support Officer, currently Dr Cecilia Di Chio, C.Di-Chio@soton.ac.uk.

Ethics Phd Application Form V1.doc Page 10



FPSE EC Application Form

Appendix (i) Participant Information template N/A

Participant Information

Ethics reference number: **ERGO/**FPSE/**xxxx**	Version: X	Date: 201y-mm-dd
Study Title: xxx		
Investigator: xxx		

Please read this information carefully before deciding to take part in this research. If you are happy to participate you may/will be / your parent / guardian will be asked to sign a consent form. Your participation is completely voluntary.

What is the research about? This is a student/research project which aims to .... The study is supported/sponsored/funded by .... At the end of the study, you will receive / may access (URL xxx) the study findings and see how your data was used.

Why have I been chosen? You have been approached because ... / You are part of a randomly selected / opportunity sample.

What will happen to me if I take part? You will first do ... and then .... It will take about ... mins in total.

Are there any benefits in my taking part? It is expected that ... / The study will add to current knowledge about .... You will receive a gift voucher / be paid for your participation.

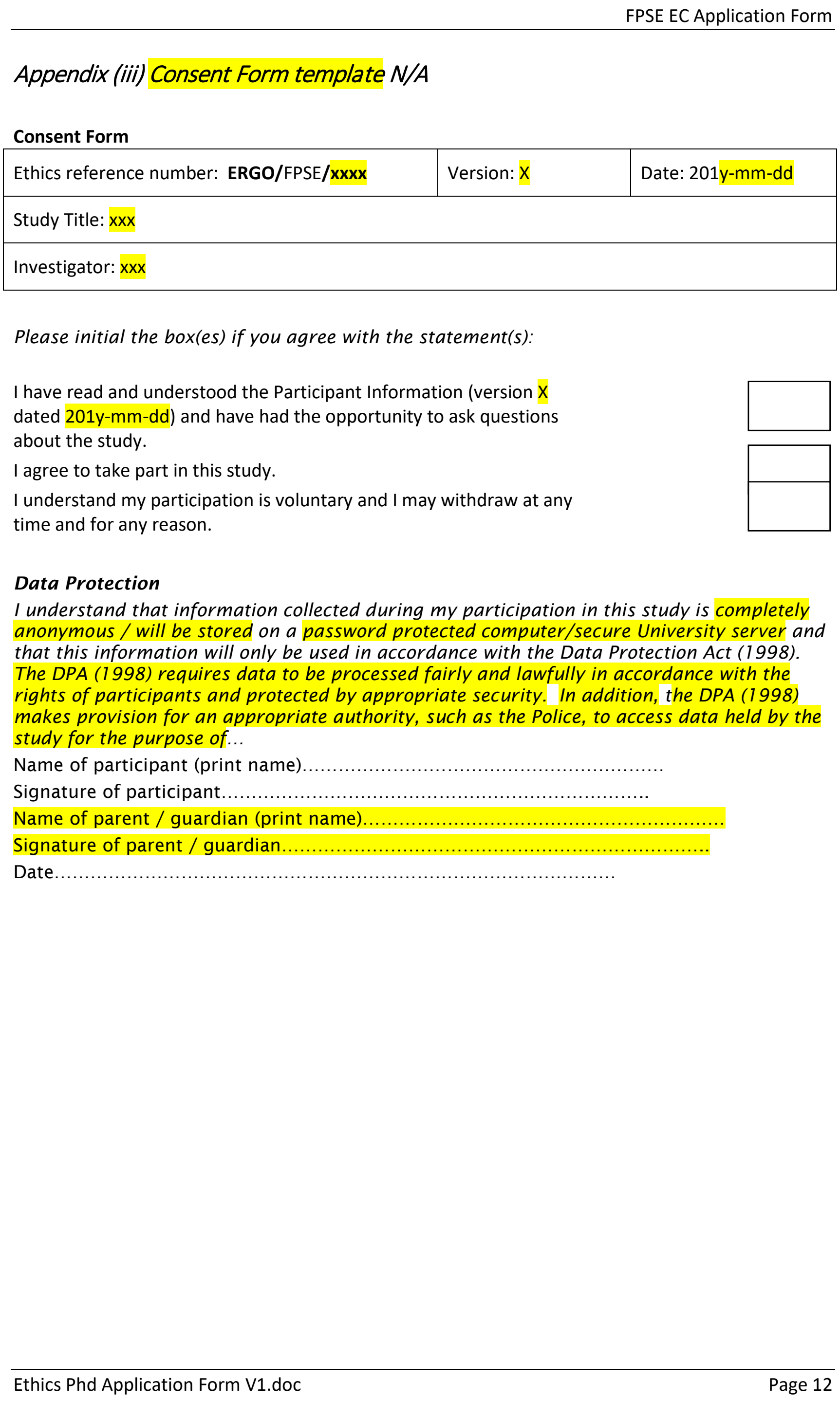
Are there any risks involved? There are no particular risks associated with your participation / There are some risks involved in .... The study will ... to minimise these risks. There will be a debrief at the end of the study, at ... on ... where you will be able to ....

Will my data be confidential? All data collected is anonymous / Your data will be held on a password protected computer/secure University server, and used only in accordance with the Data Protection Act (1998). In addition, the data will be anonymised by separating identifying data. Your data will be linked to your consent form by .... It will destroyed by .... If you would like to access your data after your participation, change it, or withdraw it, please contact the investigator (e-mail ...) or the project supervisor (e-mail ...) who will arrange this.

What happens if I change my mind? You may withdraw at any time and for any reason. You may access, change, or withdraw your data at any time and for any reason prior to its destruction. You may keep any benefits you receive.

What happens if something goes wrong? Should you have any concern or complaint, contact me if possible (investigator e-mail ...), otherwise please contact the FPSE Office (e-mail ...) or any other authoritative body such as FPSE Research Support Officer, Dr Cecilia Di Chio, C.Di-Chio@soton.ac.uk).

Ethics Phd Application Form V1.doc Page 11



Appendix (iii) Consent Form template N/A

Consent Form

Ethics reference number: **ERGO/**FPSE/**xxxx**	Version: X	Date: 201y-mm-dd
Study Title: xxx		
Investigator: xxx		

Please initial the box(es) if you agree with the statement(s):

I have read and understood the Participant Information (version X dated 201y-mm-dd) and have had the opportunity to ask questions about the study. □

I agree to take part in this study. □

I understand my participation is voluntary and I may withdraw at any time and for any reason. □

Data Protection

I understand that information collected during my participation in this study is completely anonymous / will be stored on a password protected computer/secure University server and that this information will only be used in accordance with the Data Protection Act (1998). The DPA (1998) requires data to be processed fairly and lawfully in accordance with the rights of participants and protected by appropriate security. In addition, the DPA (1998) makes provision for an appropriate authority, such as the Police, to access data held by the study for the purpose of…

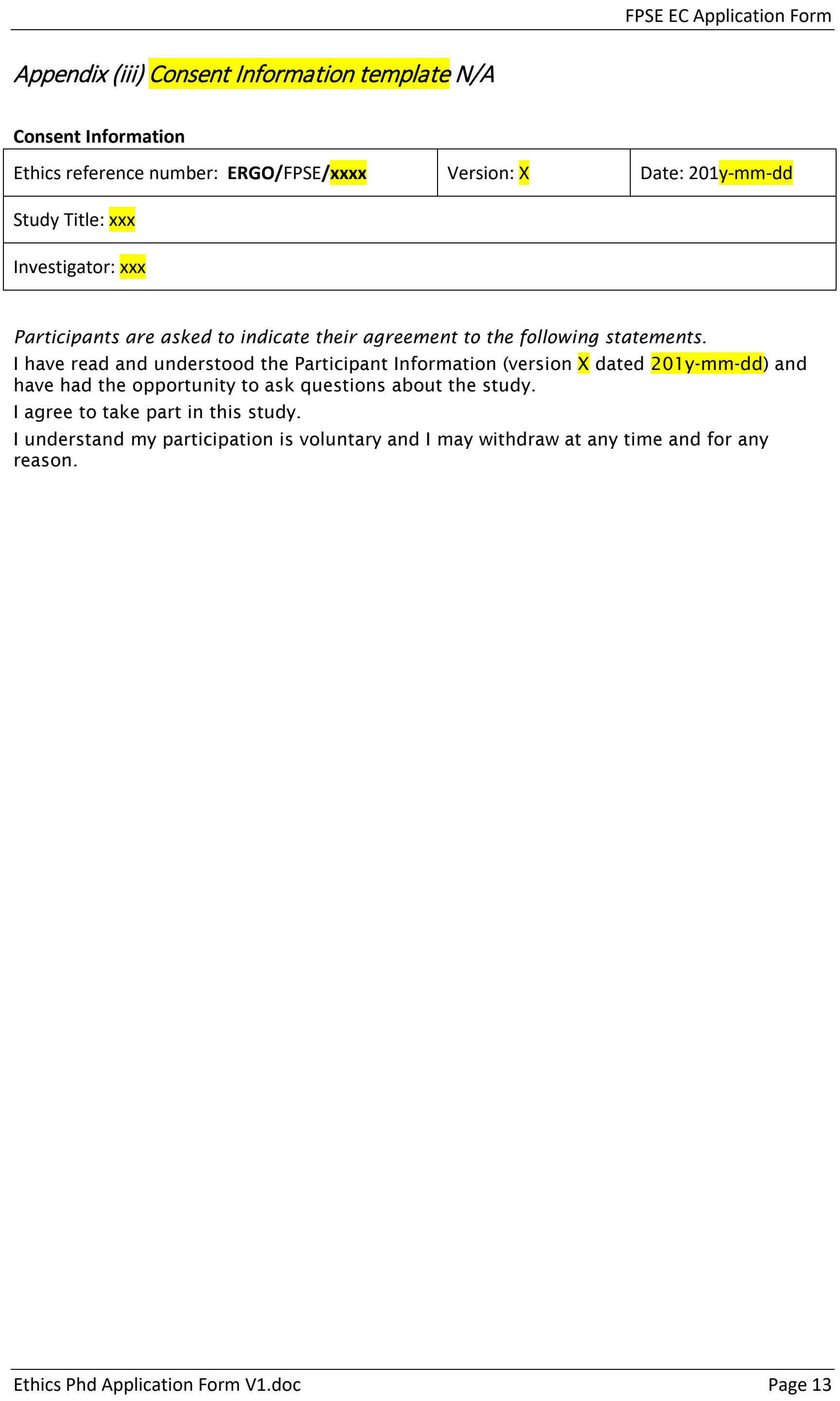
Name of participant (print name)……………………………………………………

Signature of participant……………………………………………………………….

Name of parent / guardian (print name)……………………………………………………

Signature of parent / guardian……………………………………………………………….

Date…………………………………………………………………………………



Appendix (iii) Consent Information template N/A

Consent Information

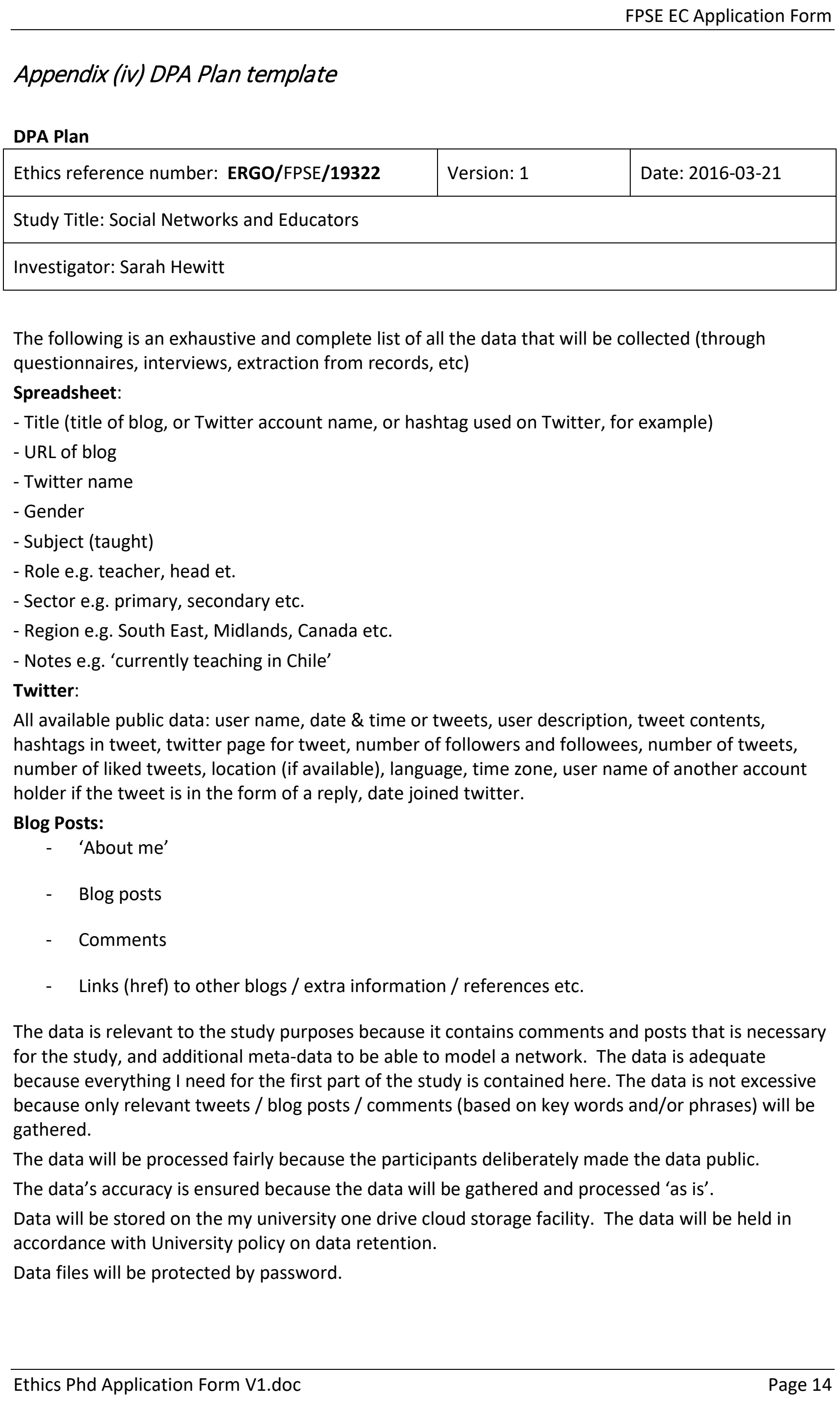
Ethics reference number: **ERGO/**FPSE**/xxxx**	Version: X	Date: 201y-mm-dd
Study Title: xxx		
Investigator: xxx		

Participants are asked to indicate their agreement to the following statements.

I have read and understood the Participant Information (version X dated 201y-mm-dd) and have had the opportunity to ask questions about the study.

I agree to take part in this study.

I understand my participation is voluntary and I may withdraw at any time and for any reason.



Appendix (iv) DPA Plan template

DPA Plan

Ethics reference number: **ERGO/**FPSE**/19322**	Version: 1	Date: 2016-03-21
Study Title: Social Networks and Educators		
Investigator: Sarah Hewitt		

The following is an exhaustive and complete list of all the data that will be collected (through questionnaires, interviews, extraction from records, etc)

Spreadsheet:

- Title (title of blog, or Twitter account name, or hashtag used on Twitter, for example)
- URL of blog
- Twitter name
- Gender
- Subject (taught)
- Role e.g. teacher, head et.
- Sector e.g. primary, secondary etc.
- Region e.g. South East, Midlands, Canada etc.
- Notes e.g. ‘currently teaching in Chile’

Twitter:

All available public data: user name, date & time or tweets, user description, tweet contents, hashtags in tweet, twitter page for tweet, number of followers and followees, number of tweets, number of liked tweets, location (if available), language, time zone, user name of another account holder if the tweet is in the form of a reply, date joined twitter.

Blog Posts:

- ‘About me’
- Blog posts
- Comments
- Links (href) to other blogs / extra information / references etc.

The data is relevant to the study purposes because it contains comments and posts that is necessary for the study, and additional meta-data to be able to model a network. The data is adequate because everything I need for the first part of the study is contained here. The data is not excessive because only relevant tweets / blog posts / comments (based on key words and/or phrases) will be gathered.

The data will be processed fairly because the participants deliberately made the data public.

The data’s accuracy is ensured because the data will be gathered and processed ‘as is’.

Data will be stored on the my university one drive cloud storage facility. The data will be held in accordance with University policy on data retention.

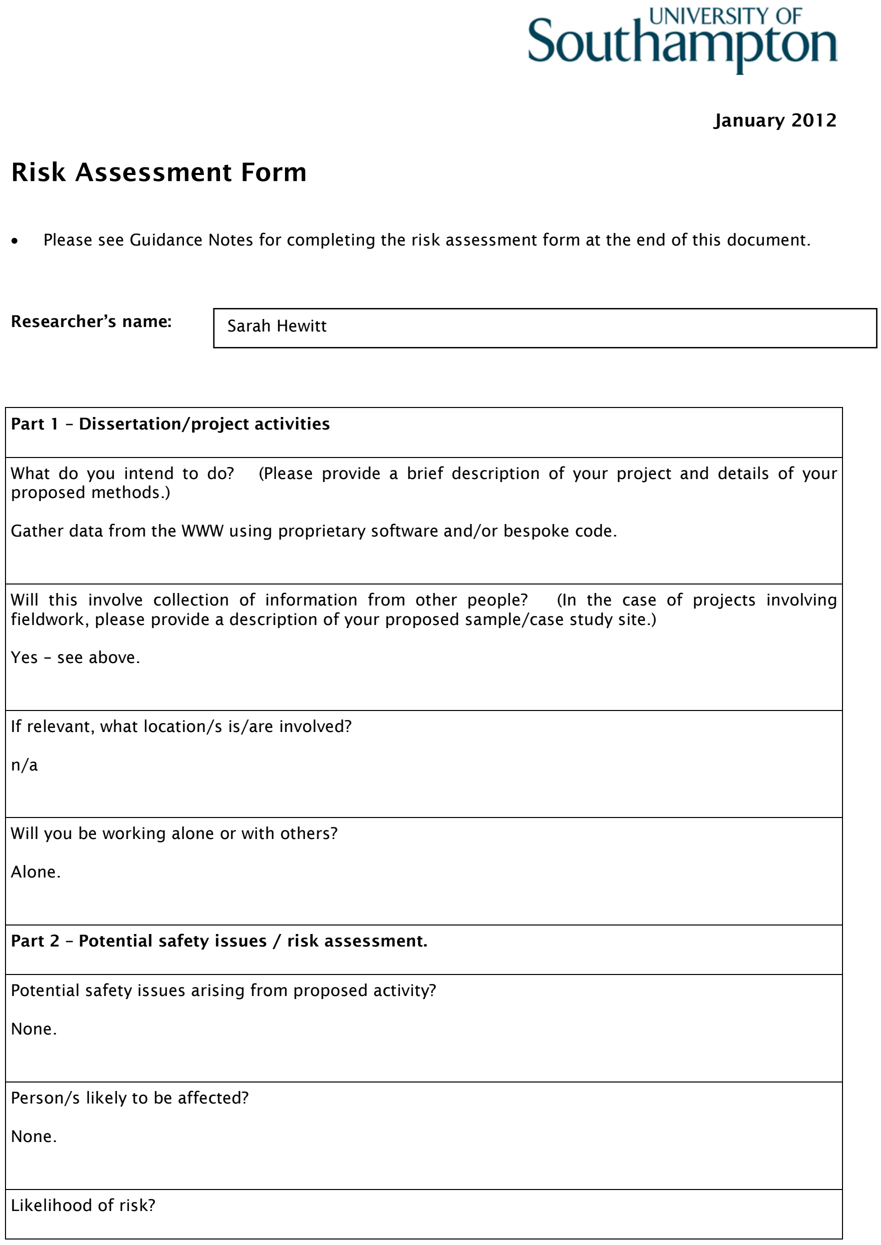
Data files will be protected by password.

The data will be destroyed by Sarah Hewitt at the University through deletion. Some of the data will be made available via the web observatory.

The data will be processed in accordance with the rights of the participants because they will have the right to access, correct, and/or withdraw their data at any time and for any reason. Participants will be able to exercise their rights by contacting the investigator (e-mail: S.Hewitt@soton.ac.uk) or the project supervisor (e-mail: T.Tiropnis@southampton.ac.uk).

The data will be anonymised by removal of meta-data. Only the content of tweets, blog posts and comments will remain.

No data will be transferred outside the European Economic Area (EEA).



UNIVERSITY OF Southampton

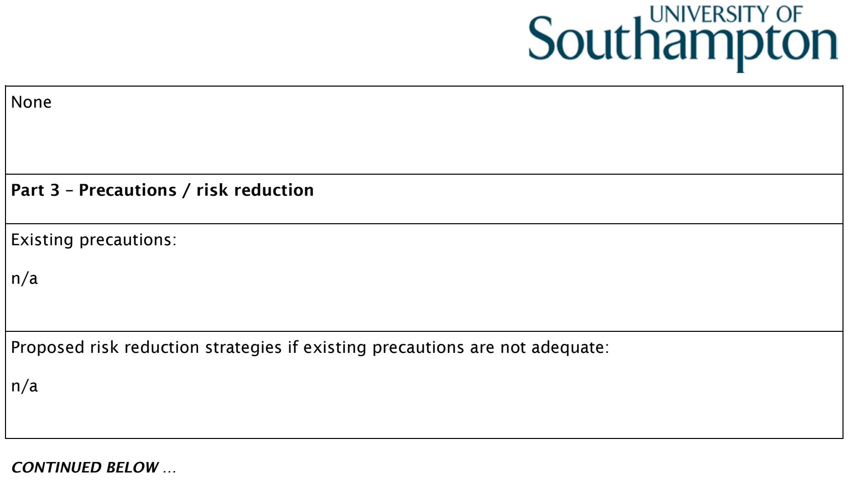
January 2012

Risk Assessment Form

- Please see Guidance Notes for completing the risk assessment form at the end of this document.

Researcher's name: Sarah Hewitt

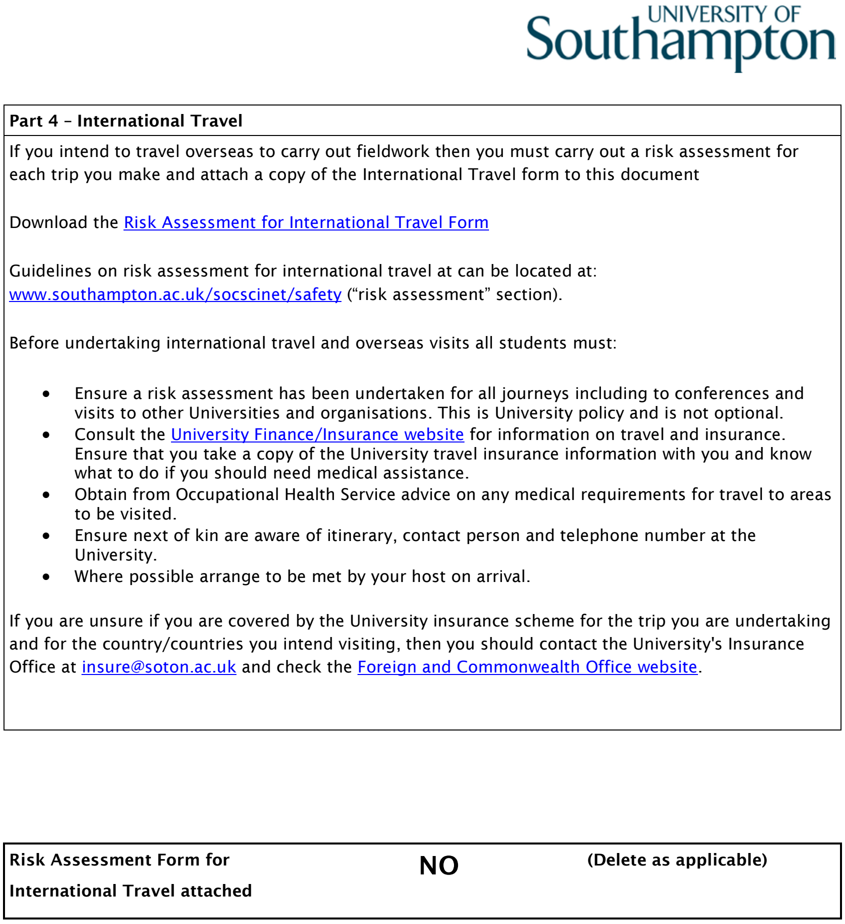
Part 1 – Dissertation/project activities
What do you intend to do? (Please provide a brief description of your project and details of your proposed methods.) Gather data from the WWW using proprietary software and/or bespoke code.
Will this involve collection of information from other people? (In the case of projects involving fieldwork, please provide a description of your proposed sample/case study site.) Yes – see above.
If relevant, what location/s is/are involved? n/a
Will you be working alone or with others? Alone.
Part 2 – Potential safety issues / risk assessment.
Potential safety issues arising from proposed activity? None.
Person/s likely to be affected? None.
Likelihood of risk?



UNIVERSITY OF Southampton

None
Part 3 – Precautions / risk reduction
Existing precautions: n/a
Proposed risk reduction strategies if existing precautions are not adequate: n/a

CONTINUED BELOW …



UNIVERSITY OF Southampton

Part 4 – International Travel

If you intend to travel overseas to carry out fieldwork then you must carry out a risk assessment for each trip you make and attach a copy of the International Travel form to this document

Download the Risk Assessment for International Travel Form

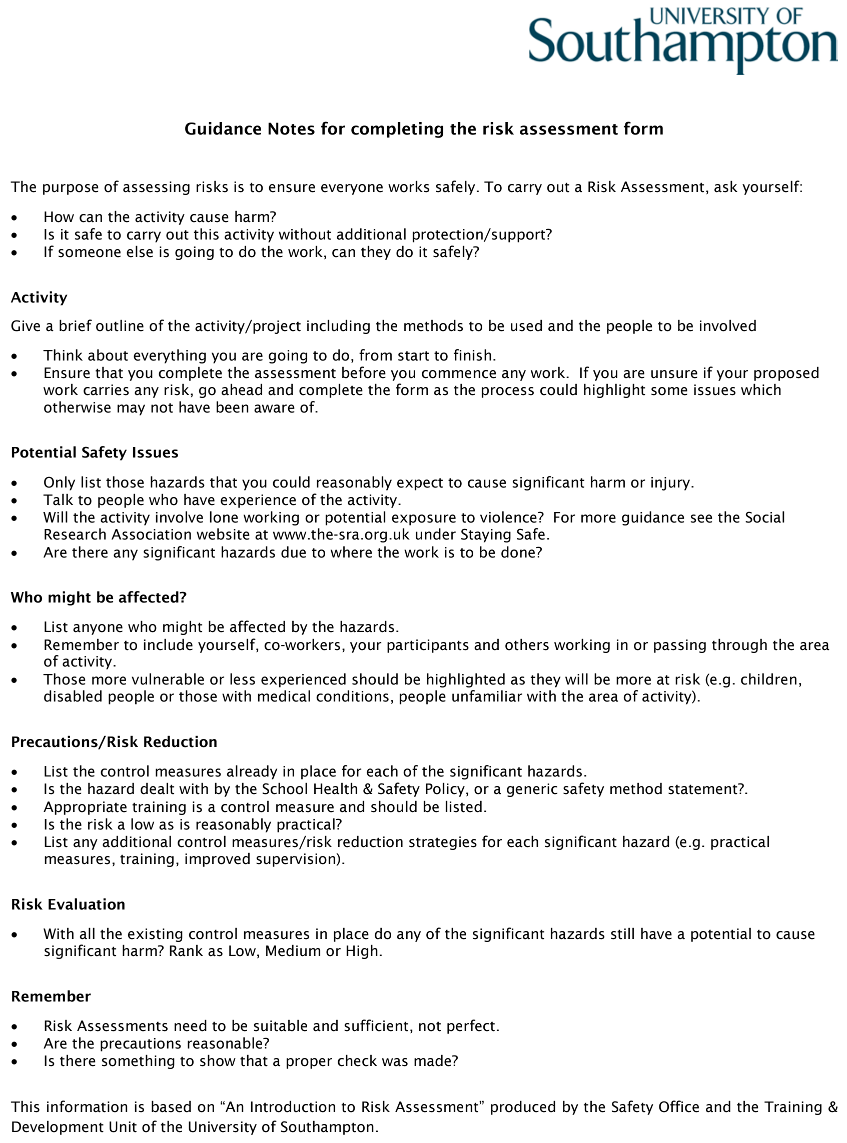
Guidelines on risk assessment for international travel at can be located at: www.southampton.ac.uk/socscinet/safety (“risk assessment” section).

Before undertaking international travel and overseas visits all students must:

- Ensure a risk assessment has been undertaken for all journeys including to conferences and visits to other Universities and organisations. This is University policy and is not optional.
- Consult the University Finance/Insurance website for information on travel and insurance. Ensure that you take a copy of the University travel insurance information with you and know what to do if you should need medical assistance.
- Obtain from Occupational Health Service advice on any medical requirements for travel to areas to be visited.
- Ensure next of kin are aware of itinerary, contact person and telephone number at the University.
- Where possible arrange to be met by your host on arrival.

If you are unsure if you are covered by the University insurance scheme for the trip you are undertaking and for the country/countries you intend visiting, then you should contact the University's Insurance Office at insure@soton.ac.uk and check the Foreign and Commonwealth Office website.

Risk Assessment Form for International Travel attached	**NO**	**(Delete as applicable)**



UNIVERSITY OF Southampton

Guidance Notes for completing the risk assessment form

The purpose of assessing risks is to ensure everyone works safely. To carry out a Risk Assessment, ask yourself:

- How can the activity cause harm?
- Is it safe to carry out this activity without additional protection/support?
- If someone else is going to do the work, can they do it safely?

Activity

Give a brief outline of the activity/project including the methods to be used and the people to be involved

- Think about everything you are going to do, from start to finish.
- Ensure that you complete the assessment before you commence any work. If you are unsure if your proposed work carries any risk, go ahead and complete the form as the process could highlight some issues which otherwise may not have been aware of.

Potential Safety Issues

- Only list those hazards that you could reasonably expect to cause significant harm or injury.
- Talk to people who have experience of the activity.
- Will the activity involve lone working or potential exposure to violence? For more guidance see the Social Research Association website at www.the-sra.org.uk under Staying Safe.
- Are there any significant hazards due to where the work is to be done?

Who might be affected?

- List anyone who might be affected by the hazards.
- Remember to include yourself, co-workers, your participants and others working in or passing through the area of activity.
- Those more vulnerable or less experienced should be highlighted as they will be more at risk (e.g. children, disabled people or those with medical conditions, people unfamiliar with the area of activity).

Precautions/Risk Reduction

- List the control measures already in place for each of the significant hazards.
- Is the hazard dealt with by the School Health & Safety Policy, or a generic safety method statement?.
- Appropriate training is a control measure and should be listed.
- Is the risk a low as is reasonably practical?
- List any additional control measures/risk reduction strategies for each significant hazard (e.g. practical measures, training, improved supervision).

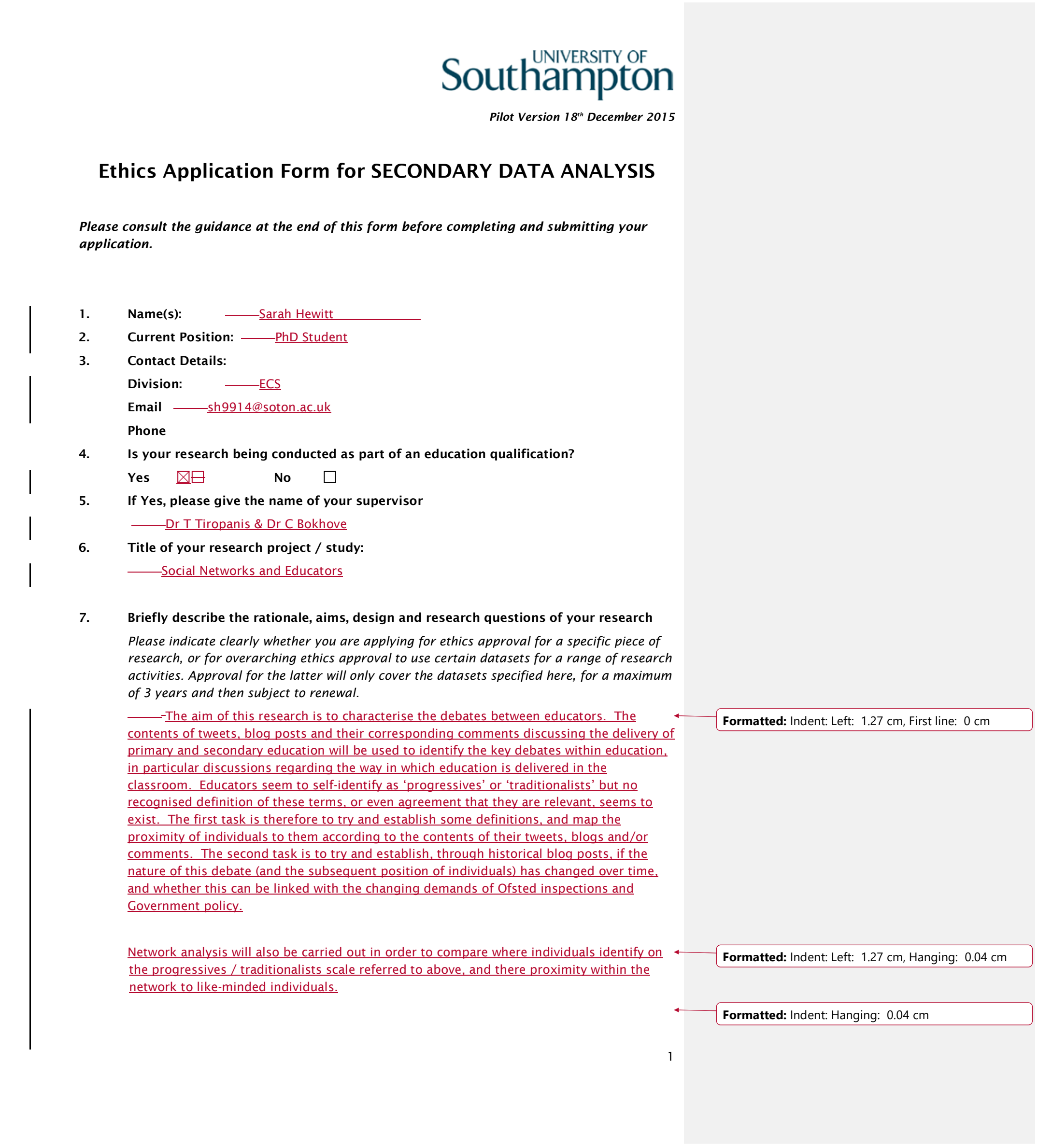
Risk Evaluation

- With all the existing control measures in place do any of the significant hazards still have a potential to cause significant harm? Rank as Low, Medium or High.

Remember

- Risk Assessments need to be suitable and sufficient, not perfect.
- Are the precautions reasonable?
- Is there something to show that a proper check was made?

This information is based on “An Introduction to Risk Assessment” produced by the Safety Office and the Training & Development Unit of the University of Southampton.



UNIVERSITY OF Southampton

Pilot Version 18th December 2015

Ethics Application Form for SECONDARY DATA ANALYSIS

Please consult the guidance at the end of this form before completing and submitting your application.

1. **Name(s):** Sarah Hewitt
2. **Current Position:** PhD Student
3. **Contact Details:**

Division: ECS

Email sh9914@soton.ac.uk

Phone

4. **Is your research being conducted as part of an education qualification?**

Yes ☒ **No** ☐

5. **If Yes, please give the name of your supervisor**

Dr T Tiropanis & Dr C Bokhove

6. **Title of your research project / study:**

Social Networks and Educators

7. **Briefly describe the rationale, aims, design and research questions of your research**

Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.

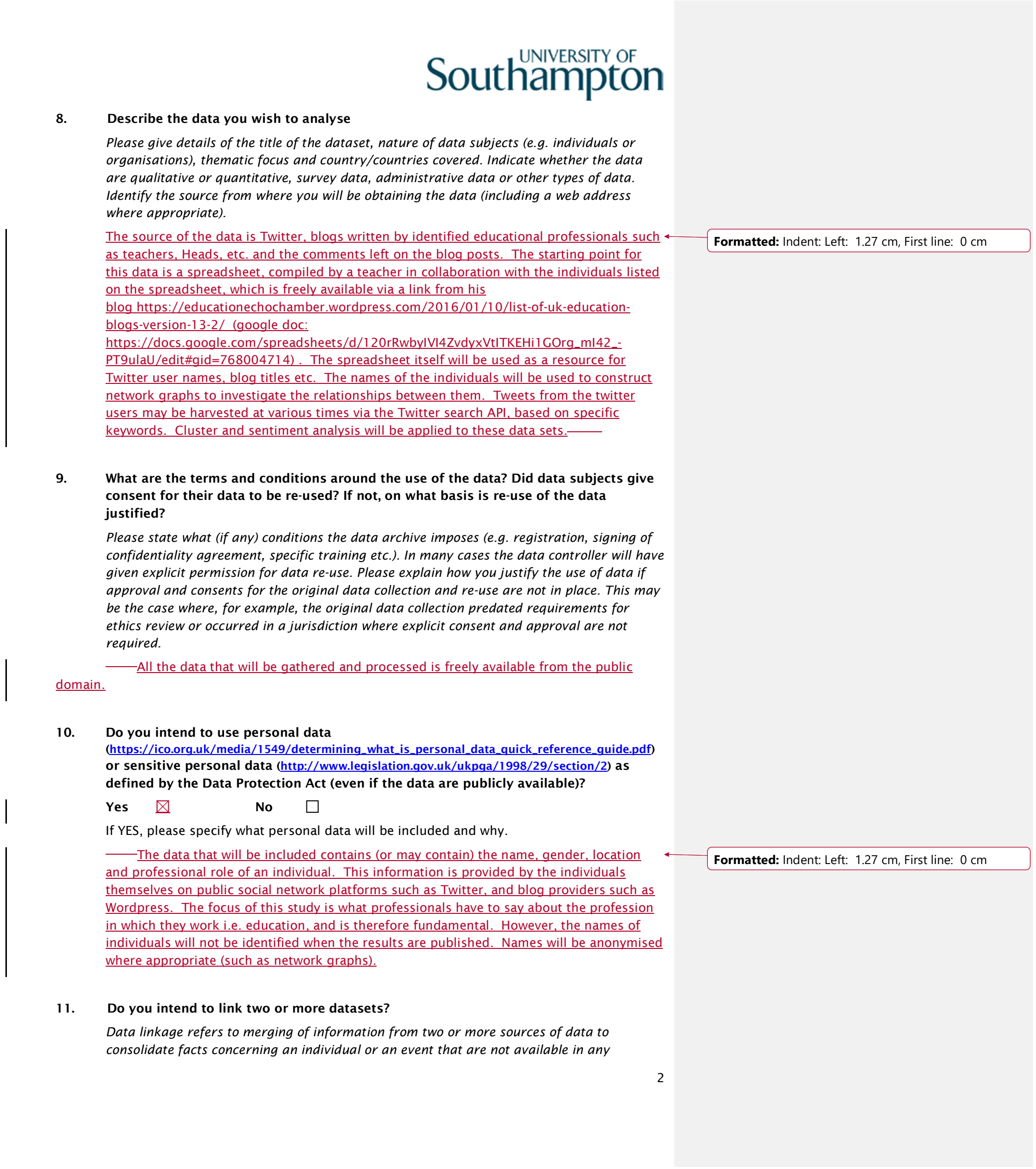
The aim of this research is to characterise the debates between educators. The contents of tweets, blog posts and their corresponding comments discussing the delivery of primary and secondary education will be used to identify the key debates within education, in particular discussions regarding the way in which education is delivered in the classroom. Educators seem to self-identify as 'progressives' or 'traditionalists' but no recognised definition of these terms, or even agreement that they are relevant, seems to exist. The first task is therefore to try and establish some definitions, and map the proximity of individuals to them according to the contents of their tweets, blogs and/or comments. The second task is to try and establish, through historical blog posts, if the nature of this debate (and the subsequent position of individuals) has changed over time, and whether this can be linked with the changing demands of Ofsted inspections and Government policy.

Formatted: Indent: Left: 1.27 cm, First line: 0 cm

Network analysis will also be carried out in order to compare where individuals identify on the progressives / traditionalists scale referred to above, and there proximity within the network to like-minded individuals.

Formatted: Indent: Left: 1.27 cm, Hanging: 0.04 cm

Formatted: Indent: Hanging: 0.04 cm



UNIVERSITY OF Southampton

8. **Describe the data you wish to analyse**

Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).

The source of the data is Twitter, blogs written by identified educational professionals such as teachers, Heads, etc. and the comments left on the blog posts. The starting point for this data is a spreadsheet, compiled by a teacher in collaboration with the individuals listed on the spreadsheet, which is freely available via a link from his blog https://educationechochamber.wordpress.com/2016/01/10/list-of-uk-education-blogs-version-13-2/ (google doc: https://docs.google.com/spreadsheets/d/120rRwbyIVI4ZvdyxVtITKEHi1GOrg_mI42_-PT9ulaU/edit#gid=768004714) . The spreadsheet itself will be used as a resource for Twitter user names, blog titles etc. The names of the individuals will be used to construct network graphs to investigate the relationships between them. Tweets from the twitter users may be harvested at various times via the Twitter search API, based on specific keywords. Cluster and sentiment analysis will be applied to these data sets.

Formatted: Indent: Left: 1.27 cm, First line: 0 cm

9. **What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?**

Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.

All the data that will be gathered and processed is freely available from the public domain.

10. **Do you intend to use personal data (https://ico.org.uk/media/1549/determining_what_is_personal_data_quick_reference_guide.pdf) or sensitive personal data (http://www.legislation.gov.uk/ukpga/1998/29/section/2) as defined by the Data Protection Act (even if the data are publicly available)?**

Yes ☒ **No** □

If YES, please specify what personal data will be included and why.

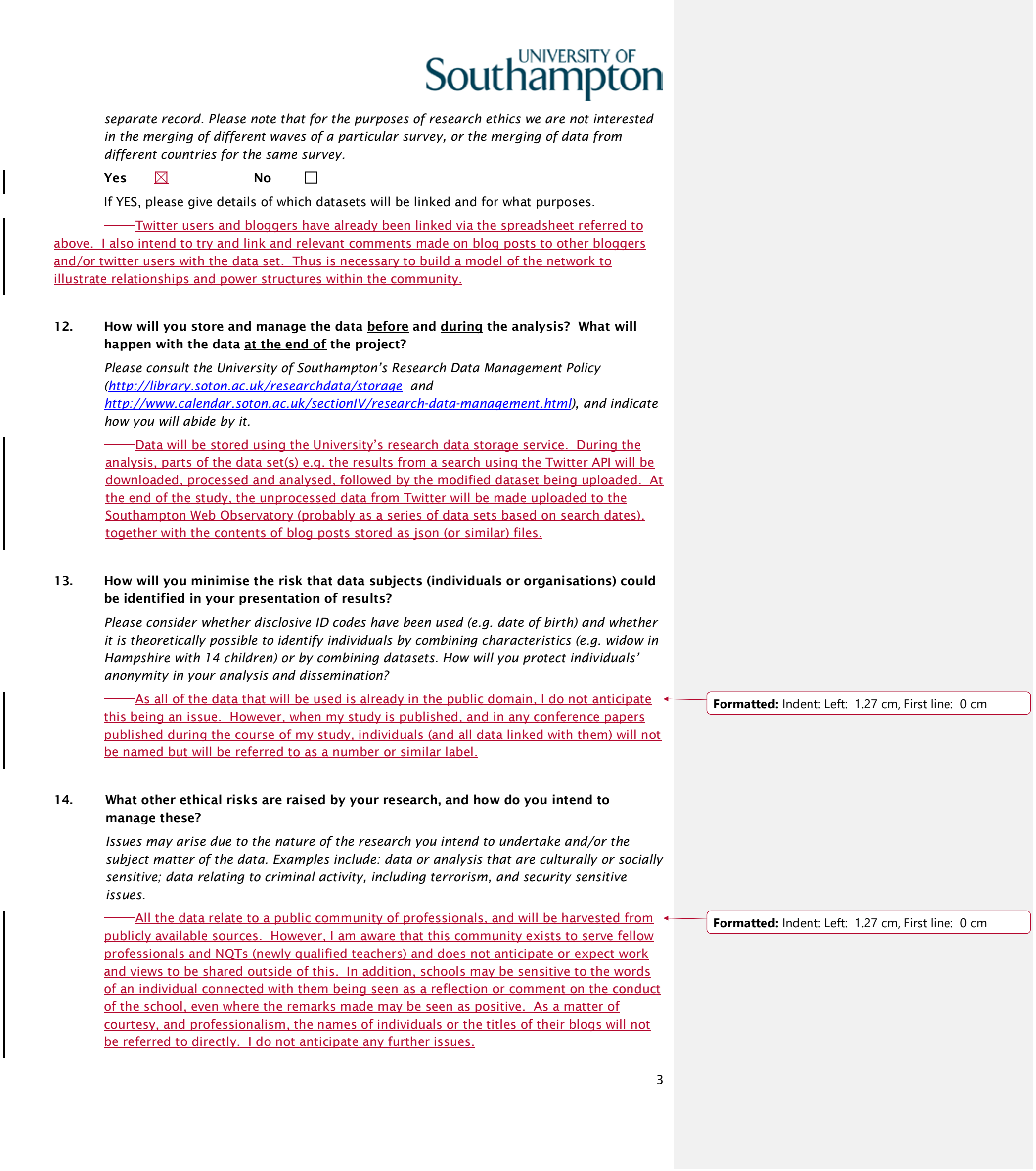
The data that will be included contains (or may contain) the name, gender, location and professional role of an individual. This information is provided by the individuals themselves on public social network platforms such as Twitter, and blog providers such as Wordpress. The focus of this study is what professionals have to say about the profession in which they work i.e. education, and is therefore fundamental. However, the names of individuals will not be identified when the results are published. Names will be anonymised where appropriate (such as network graphs).

Formatted: Indent: Left: 1.27 cm, First line: 0 cm

11. **Do you intend to link two or more datasets?**

Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any

2



UNIVERSITY OF Southampton

separate record. Please note that for the purposes of research ethics we are not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.

Yes ☒ **No** □

If YES, please give details of which datasets will be linked and for what purposes.

Twitter users and bloggers have already been linked via the spreadsheet referred to above. I also intend to try and link and relevant comments made on blog posts to other bloggers and/or twitter users with the data set. Thus is necessary to build a model of the network to illustrate relationships and power structures within the community.

12. How will you store and manage the data before and during the analysis? What will happen with the data at the end of the project?

Please consult the University of Southampton's Research Data Management Policy (http://library.soton.ac.uk/researchdata/storage and http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html), and indicate how you will abide by it.

Data will be stored using the University's research data storage service. During the analysis, parts of the data set(s) e.g. the results from a search using the Twitter API will be downloaded, processed and analysed, followed by the modified dataset being uploaded. At the end of the study, the unprocessed data from Twitter will be made uploaded to the Southampton Web Observatory (probably as a series of data sets based on search dates), together with the contents of blog posts stored as json (or similar) files.

13. How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?

Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?

As all of the data that will be used is already in the public domain, I do not anticipate this being an issue. However, when my study is published, and in any conference papers published during the course of my study, individuals (and all data linked with them) will not be named but will be referred to as a number or similar label.

Formatted: Indent: Left: 1.27 cm, First line: 0 cm

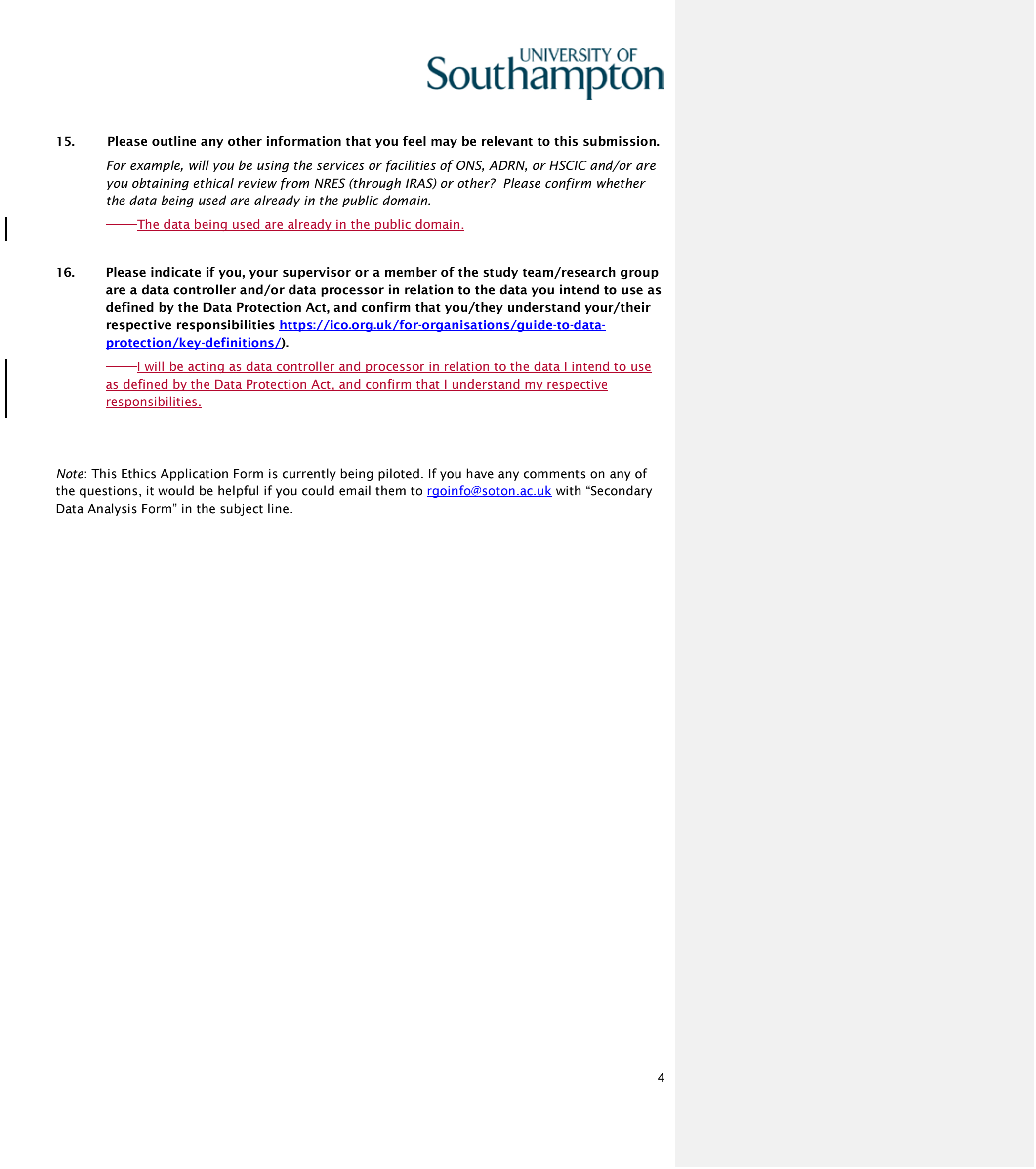
14. What other ethical risks are raised by your research, and how do you intend to manage these?

Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.

All the data relate to a public community of professionals, and will be harvested from publicly available sources. However, I am aware that this community exists to serve fellow professionals and NQTs (newly qualified teachers) and does not anticipate or expect work and views to be shared outside of this. In addition, schools may be sensitive to the words of an individual connected with them being seen as a reflection or comment on the conduct of the school, even where the remarks made may be seen as positive. As a matter of courtesy, and professionalism, the names of individuals or the titles of their blogs will not be referred to directly. I do not anticipate any further issues.

Formatted: Indent: Left: 1.27 cm, First line: 0 cm

3



UNIVERSITY OF Southampton

15. **Please outline any other information that you feel may be relevant to this submission.**

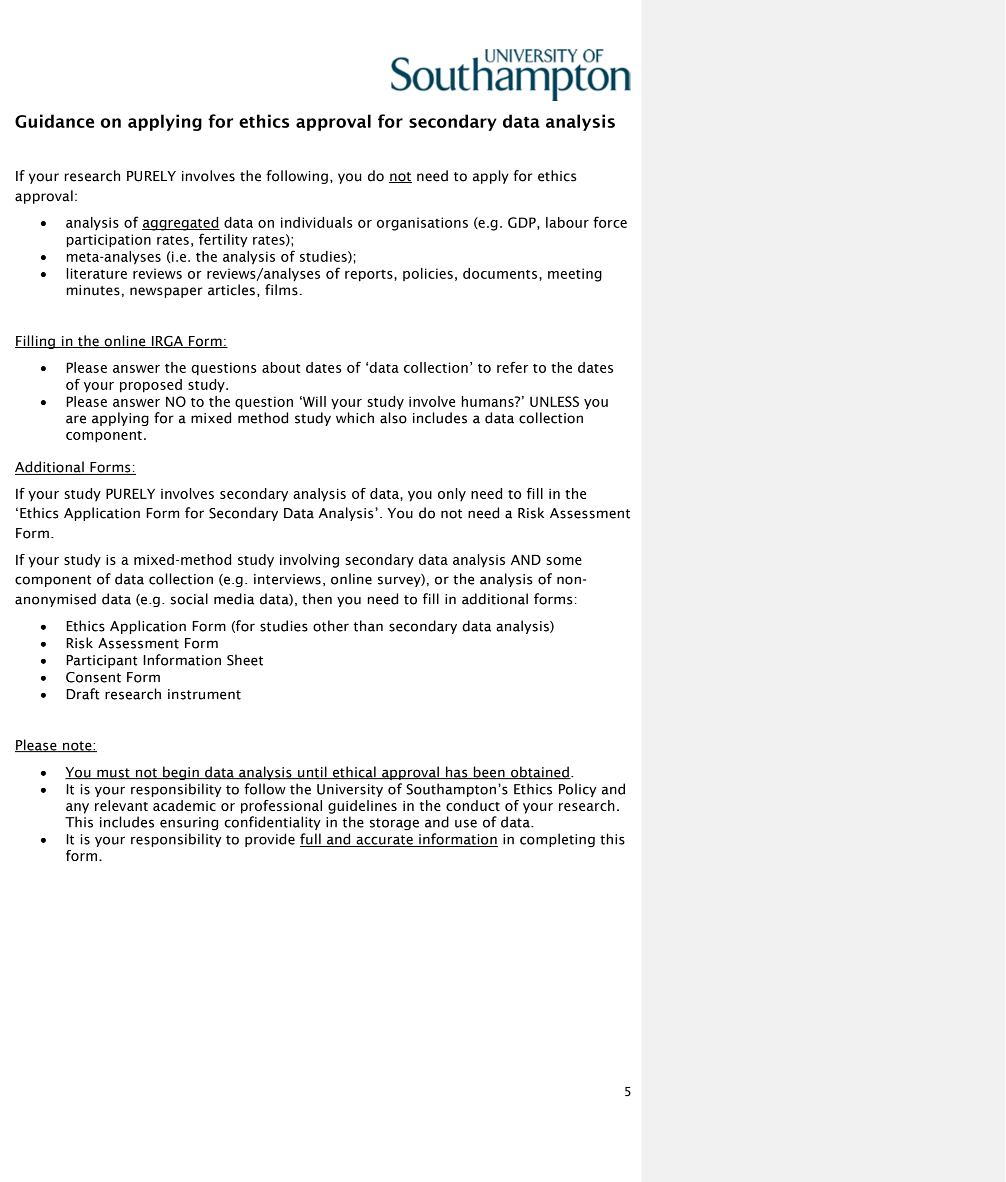
For example, will you be using the services or facilities of ONS, ADRN, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.

The data being used are already in the public domain.

16. **Please indicate if you, your supervisor or a member of the study team/research group are a data controller and/or data processor in relation to the data you intend to use as defined by the Data Protection Act, and confirm that you/they understand your/their respective responsibilities https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/).**

I will be acting as data controller and processor in relation to the data I intend to use as defined by the Data Protection Act, and confirm that I understand my respective responsibilities.

Note: This Ethics Application Form is currently being piloted. If you have any comments on any of the questions, it would be helpful if you could email them to rgoinfo@soton.ac.uk with “Secondary Data Analysis Form” in the subject line.



UNIVERSITY OF Southampton

Guidance on applying for ethics approval for secondary data analysis

If your research PURELY involves the following, you do not need to apply for ethics approval:

- analysis of aggregated data on individuals or organisations (e.g. GDP, labour force participation rates, fertility rates);
- meta-analyses (i.e. the analysis of studies);
- literature reviews or reviews/analyses of reports, policies, documents, meeting minutes, newspaper articles, films.

Filling in the online IRGA Form:

- Please answer the questions about dates of 'data collection' to refer to the dates of your proposed study.
- Please answer NO to the question 'Will your study involve humans?' UNLESS you are applying for a mixed method study which also includes a data collection component.

Additional Forms:

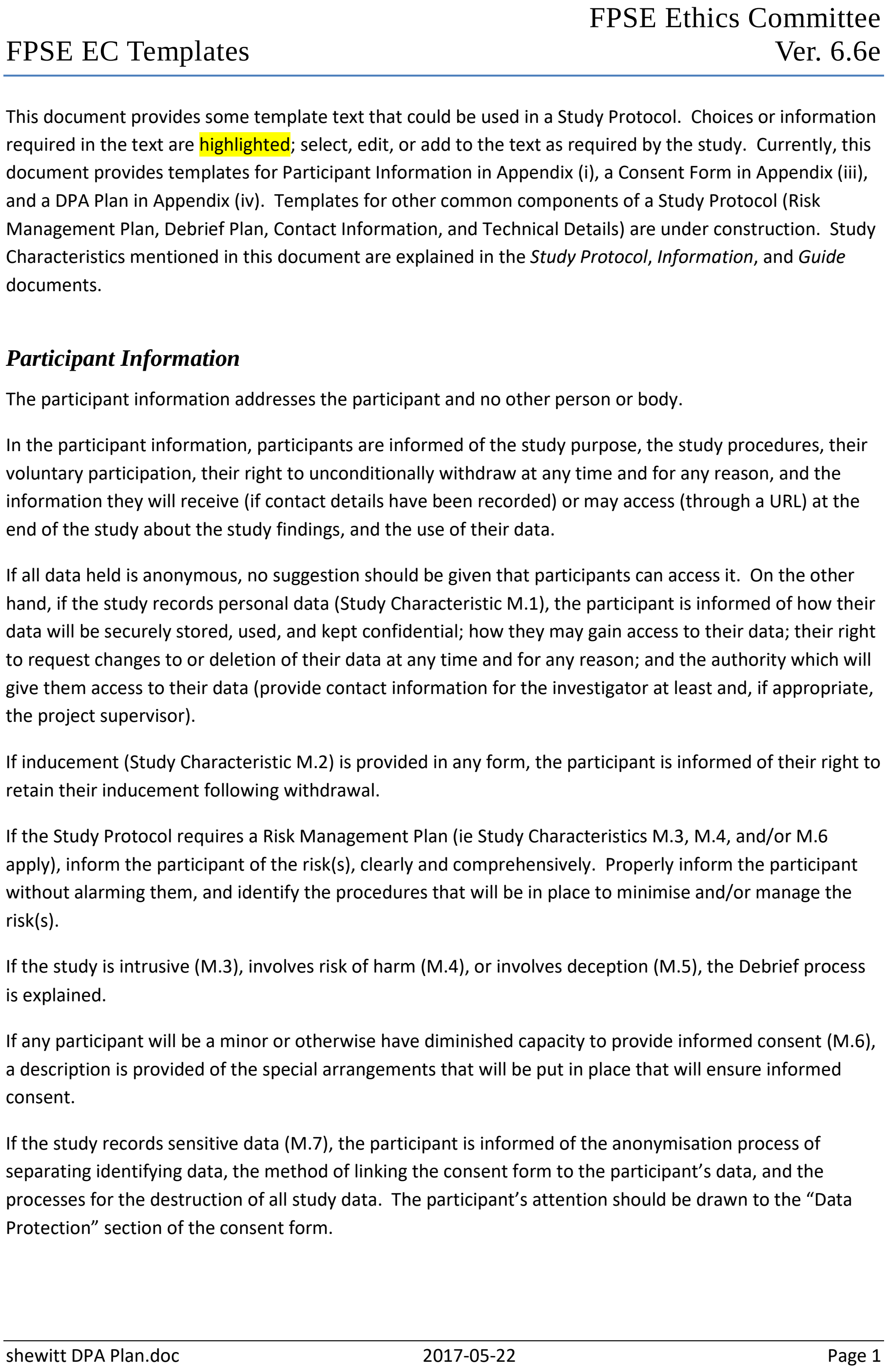
If your study PURELY involves secondary analysis of data, you only need to fill in the 'Ethics Application Form for Secondary Data Analysis'. You do not need a Risk Assessment Form.

If your study is a mixed-method study involving secondary data analysis AND some component of data collection (e.g. interviews, online survey), or the analysis of non-anonymised data (e.g. social media data), then you need to fill in additional forms:

- Ethics Application Form (for studies other than secondary data analysis)
- Risk Assessment Form
- Participant Information Sheet
- Consent Form
- Draft research instrument

Please note:

- You must not begin data analysis until ethical approval has been obtained.
- It is your responsibility to follow the University of Southampton's Ethics Policy and any relevant academic or professional guidelines in the conduct of your research. This includes ensuring confidentiality in the storage and use of data.
- It is your responsibility to provide full and accurate information in completing this form.



FPSE Ethics Committee

FPSE EC Templates Ver. 6.6e

This document provides some template text that could be used in a Study Protocol. Choices or information required in the text are highlighted; select, edit, or add to the text as required by the study. Currently, this document provides templates for Participant Information in Appendix (i), a Consent Form in Appendix (iii), and a DPA Plan in Appendix (iv). Templates for other common components of a Study Protocol (Risk Management Plan, Debrief Plan, Contact Information, and Technical Details) are under construction. Study Characteristics mentioned in this document are explained in the *Study Protocol*, *Information*, and *Guide* documents.

Participant Information

The participant information addresses the participant and no other person or body.

In the participant information, participants are informed of the study purpose, the study procedures, their voluntary participation, their right to unconditionally withdraw at any time and for any reason, and the information they will receive (if contact details have been recorded) or may access (through a URL) at the end of the study about the study findings, and the use of their data.

If all data held is anonymous, no suggestion should be given that participants can access it. On the other hand, if the study records personal data (Study Characteristic M.1), the participant is informed of how their data will be securely stored, used, and kept confidential; how they may gain access to their data; their right to request changes to or deletion of their data at any time and for any reason; and the authority which will give them access to their data (provide contact information for the investigator at least and, if appropriate, the project supervisor).

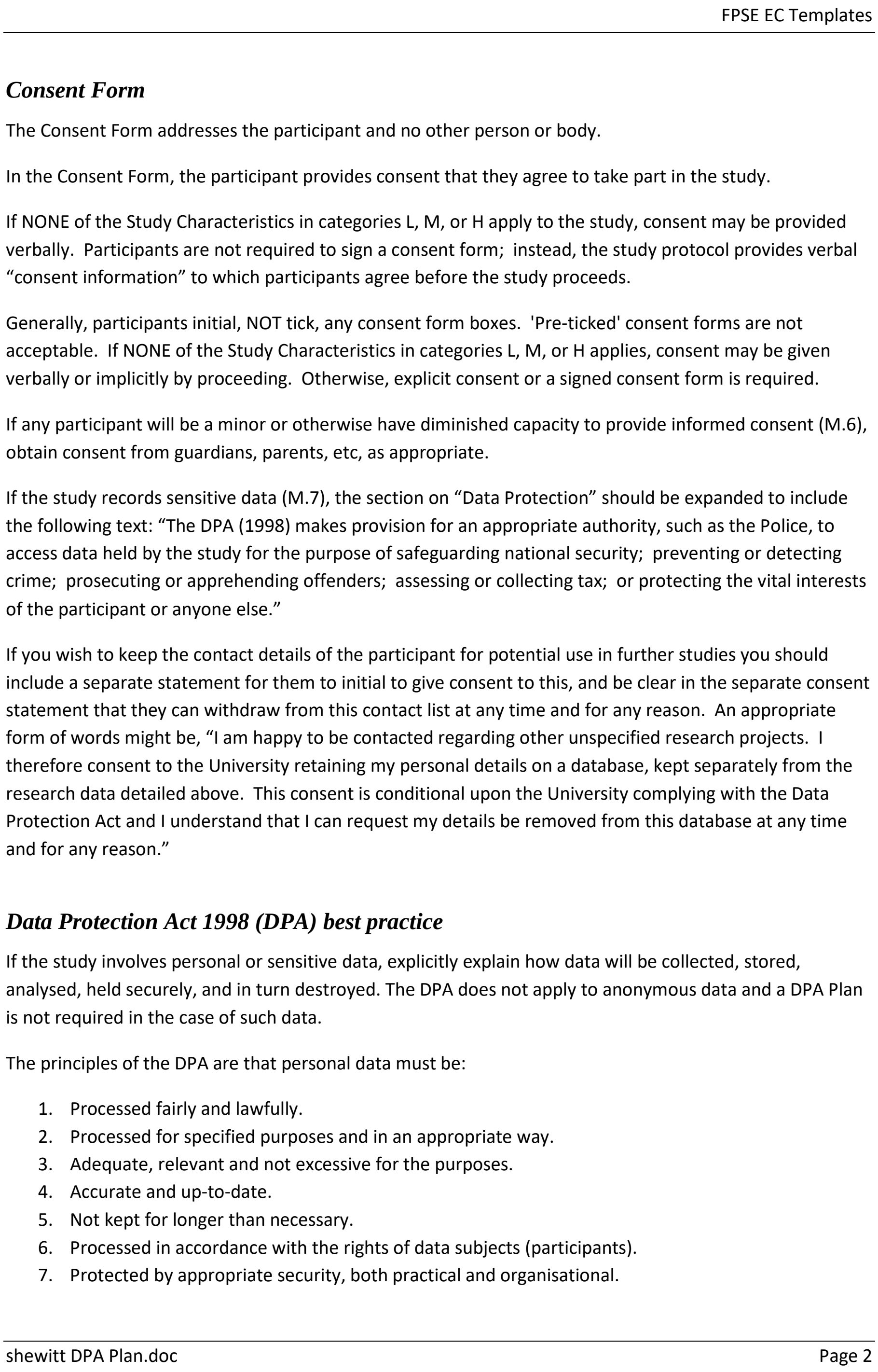
If inducement (Study Characteristic M.2) is provided in any form, the participant is informed of their right to retain their inducement following withdrawal.

If the Study Protocol requires a Risk Management Plan (ie Study Characteristics M.3, M.4, and/or M.6 apply), inform the participant of the risk(s), clearly and comprehensively. Properly inform the participant without alarming them, and identify the procedures that will be in place to minimise and/or manage the risk(s).

If the study is intrusive (M.3), involves risk of harm (M.4), or involves deception (M.5), the Debrief process is explained.

If any participant will be a minor or otherwise have diminished capacity to provide informed consent (M.6), a description is provided of the special arrangements that will be put in place that will ensure informed consent.

If the study records sensitive data (M.7), the participant is informed of the anonymisation process of separating identifying data, the method of linking the consent form to the participant's data, and the processes for the destruction of all study data. The participant's attention should be drawn to the "Data Protection" section of the consent form.



Consent Form

The Consent Form addresses the participant and no other person or body.

In the Consent Form, the participant provides consent that they agree to take part in the study.

If NONE of the Study Characteristics in categories L, M, or H apply to the study, consent may be provided verbally. Participants are not required to sign a consent form; instead, the study protocol provides verbal “consent information” to which participants agree before the study proceeds.

Generally, participants initial, NOT tick, any consent form boxes. 'Pre-ticked' consent forms are not acceptable. If NONE of the Study Characteristics in categories L, M, or H applies, consent may be given verbally or implicitly by proceeding. Otherwise, explicit consent or a signed consent form is required.

If any participant will be a minor or otherwise have diminished capacity to provide informed consent (M.6), obtain consent from guardians, parents, etc, as appropriate.

If the study records sensitive data (M.7), the section on “Data Protection” should be expanded to include the following text: “The DPA (1998) makes provision for an appropriate authority, such as the Police, to access data held by the study for the purpose of safeguarding national security; preventing or detecting crime; prosecuting or apprehending offenders; assessing or collecting tax; or protecting the vital interests of the participant or anyone else.”

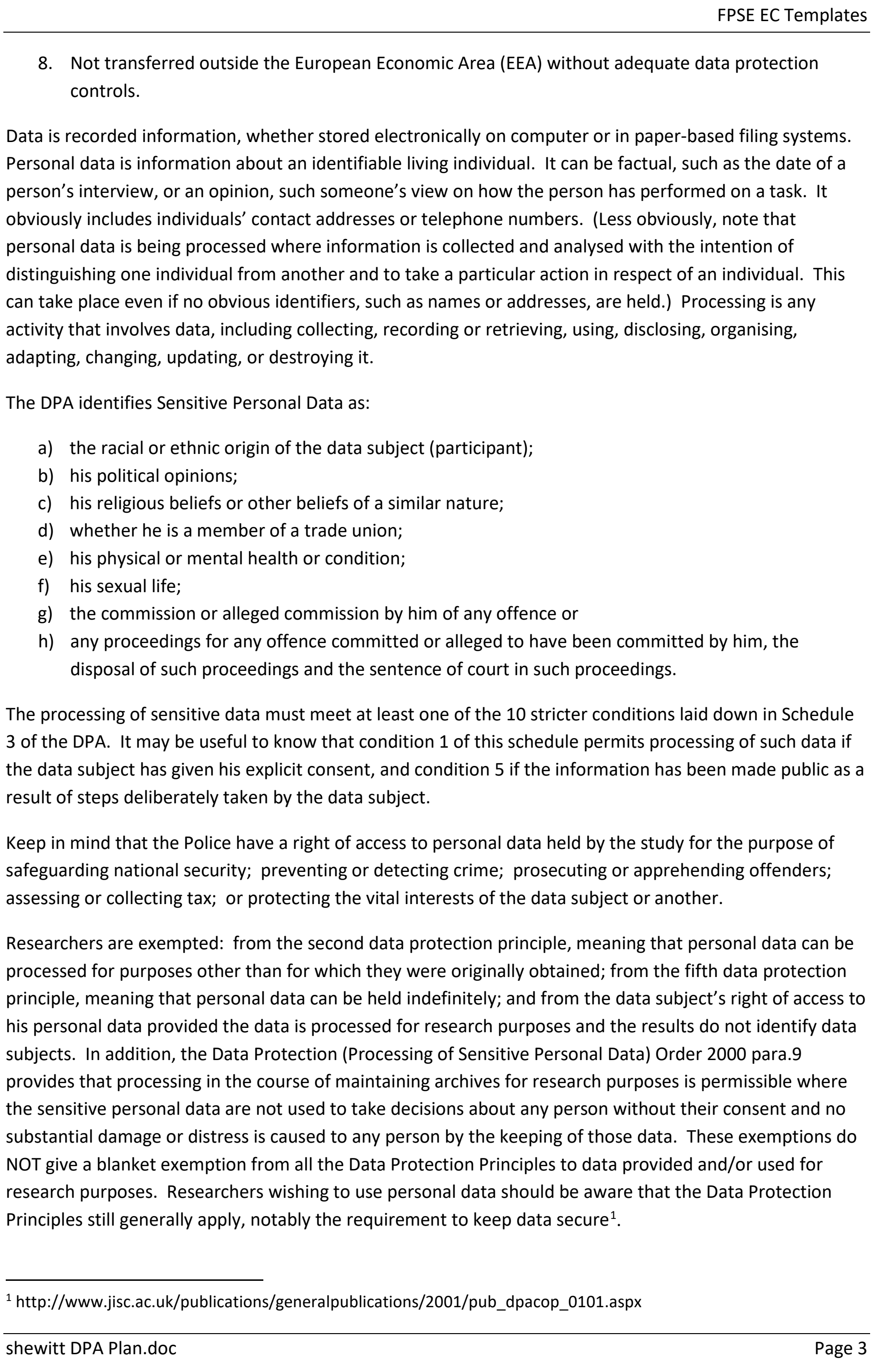
If you wish to keep the contact details of the participant for potential use in further studies you should include a separate statement for them to initial to give consent to this, and be clear in the separate consent statement that they can withdraw from this contact list at any time and for any reason. An appropriate form of words might be, “I am happy to be contacted regarding other unspecified research projects. I therefore consent to the University retaining my personal details on a database, kept separately from the research data detailed above. This consent is conditional upon the University complying with the Data Protection Act and I understand that I can request my details be removed from this database at any time and for any reason.”

Data Protection Act 1998 (DPA) best practice

If the study involves personal or sensitive data, explicitly explain how data will be collected, stored, analysed, held securely, and in turn destroyed. The DPA does not apply to anonymous data and a DPA Plan is not required in the case of such data.

The principles of the DPA are that personal data must be:

1. Processed fairly and lawfully.
2. Processed for specified purposes and in an appropriate way.
3. Adequate, relevant and not excessive for the purposes.
4. Accurate and up-to-date.
5. Not kept for longer than necessary.
6. Processed in accordance with the rights of data subjects (participants).
7. Protected by appropriate security, both practical and organisational.



8. Not transferred outside the European Economic Area (EEA) without adequate data protection controls.

Data is recorded information, whether stored electronically on computer or in paper-based filing systems. Personal data is information about an identifiable living individual. It can be factual, such as the date of a person's interview, or an opinion, such someone's view on how the person has performed on a task. It obviously includes individuals' contact addresses or telephone numbers. (Less obviously, note that personal data is being processed where information is collected and analysed with the intention of distinguishing one individual from another and to take a particular action in respect of an individual. This can take place even if no obvious identifiers, such as names or addresses, are held.) Processing is any activity that involves data, including collecting, recording or retrieving, using, disclosing, organising, adapting, changing, updating, or destroying it.

The DPA identifies Sensitive Personal Data as:

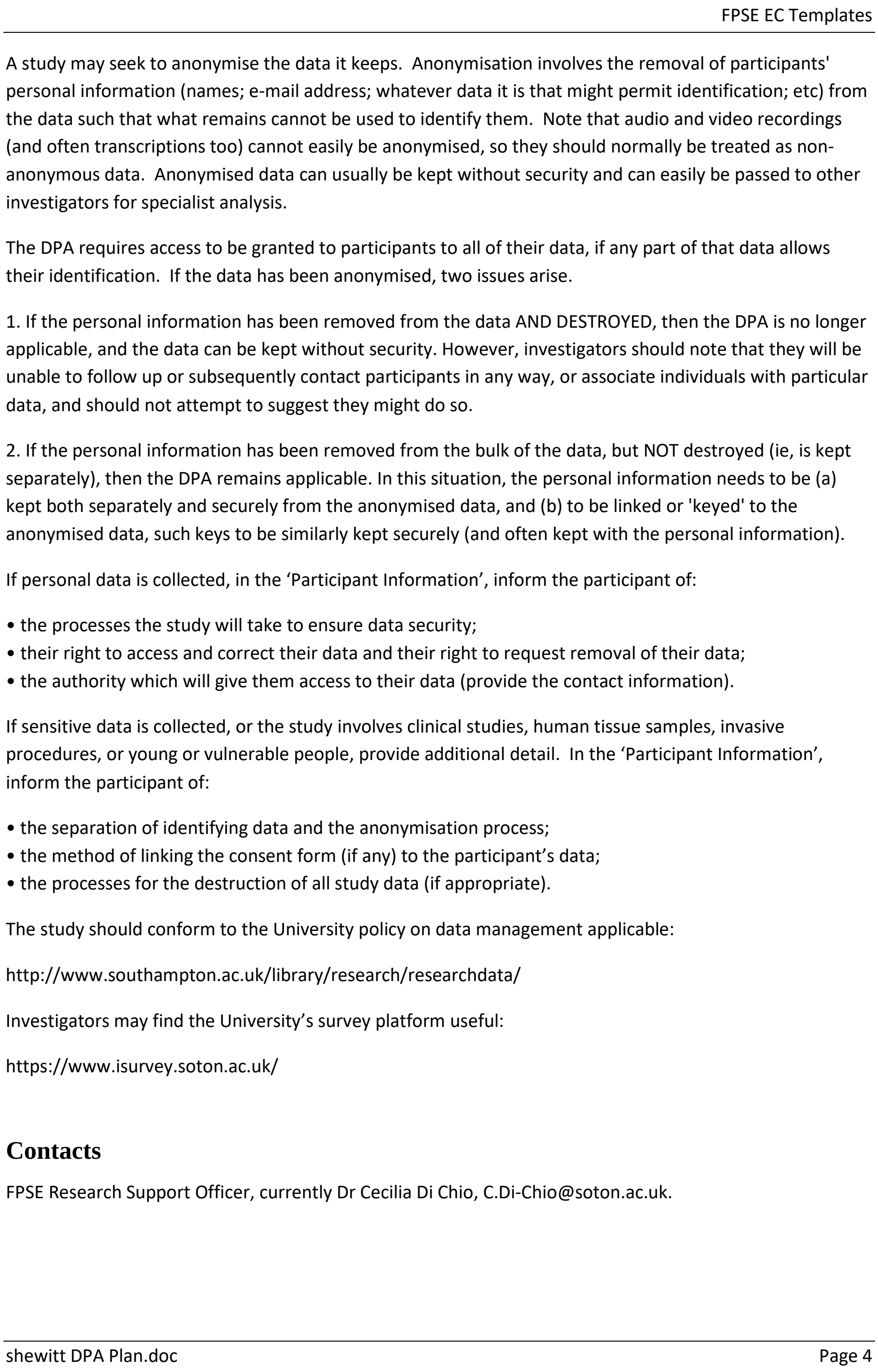
a) the racial or ethnic origin of the data subject (participant);
b) his political opinions;
c) his religious beliefs or other beliefs of a similar nature;
d) whether he is a member of a trade union;
e) his physical or mental health or condition;
f) his sexual life;
g) the commission or alleged commission by him of any offence or
h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings and the sentence of court in such proceedings.

The processing of sensitive data must meet at least one of the 10 stricter conditions laid down in Schedule 3 of the DPA. It may be useful to know that condition 1 of this schedule permits processing of such data if the data subject has given his explicit consent, and condition 5 if the information has been made public as a result of steps deliberately taken by the data subject.

Keep in mind that the Police have a right of access to personal data held by the study for the purpose of safeguarding national security; preventing or detecting crime; prosecuting or apprehending offenders; assessing or collecting tax; or protecting the vital interests of the data subject or another.

Researchers are exempted: from the second data protection principle, meaning that personal data can be processed for purposes other than for which they were originally obtained; from the fifth data protection principle, meaning that personal data can be held indefinitely; and from the data subject's right of access to his personal data provided the data is processed for research purposes and the results do not identify data subjects. In addition, the Data Protection (Processing of Sensitive Personal Data) Order 2000 para.9 provides that processing in the course of maintaining archives for research purposes is permissible where the sensitive personal data are not used to take decisions about any person without their consent and no substantial damage or distress is caused to any person by the keeping of those data. These exemptions do NOT give a blanket exemption from all the Data Protection Principles to data provided and/or used for research purposes. Researchers wishing to use personal data should be aware that the Data Protection Principles still generally apply, notably the requirement to keep data secure[1].

[1] http://www.jisc.ac.uk/publications/generalpublications/2001/pub_dpacop_0101.aspx



A study may seek to anonymise the data it keeps. Anonymisation involves the removal of participants' personal information (names; e-mail address; whatever data it is that might permit identification; etc) from the data such that what remains cannot be used to identify them. Note that audio and video recordings (and often transcriptions too) cannot easily be anonymised, so they should normally be treated as non-anonymous data. Anonymised data can usually be kept without security and can easily be passed to other investigators for specialist analysis.

The DPA requires access to be granted to participants to all of their data, if any part of that data allows their identification. If the data has been anonymised, two issues arise.

1. If the personal information has been removed from the data AND DESTROYED, then the DPA is no longer applicable, and the data can be kept without security. However, investigators should note that they will be unable to follow up or subsequently contact participants in any way, or associate individuals with particular data, and should not attempt to suggest they might do so.

2. If the personal information has been removed from the bulk of the data, but NOT destroyed (ie, is kept separately), then the DPA remains applicable. In this situation, the personal information needs to be (a) kept both separately and securely from the anonymised data, and (b) to be linked or 'keyed' to the anonymised data, such keys to be similarly kept securely (and often kept with the personal information).

If personal data is collected, in the ‘Participant Information’, inform the participant of:

- the processes the study will take to ensure data security;
- their right to access and correct their data and their right to request removal of their data;
- the authority which will give them access to their data (provide the contact information).

If sensitive data is collected, or the study involves clinical studies, human tissue samples, invasive procedures, or young or vulnerable people, provide additional detail. In the ‘Participant Information’, inform the participant of:

- the separation of identifying data and the anonymisation process;
- the method of linking the consent form (if any) to the participant’s data;
- the processes for the destruction of all study data (if appropriate).

The study should conform to the University policy on data management applicable:

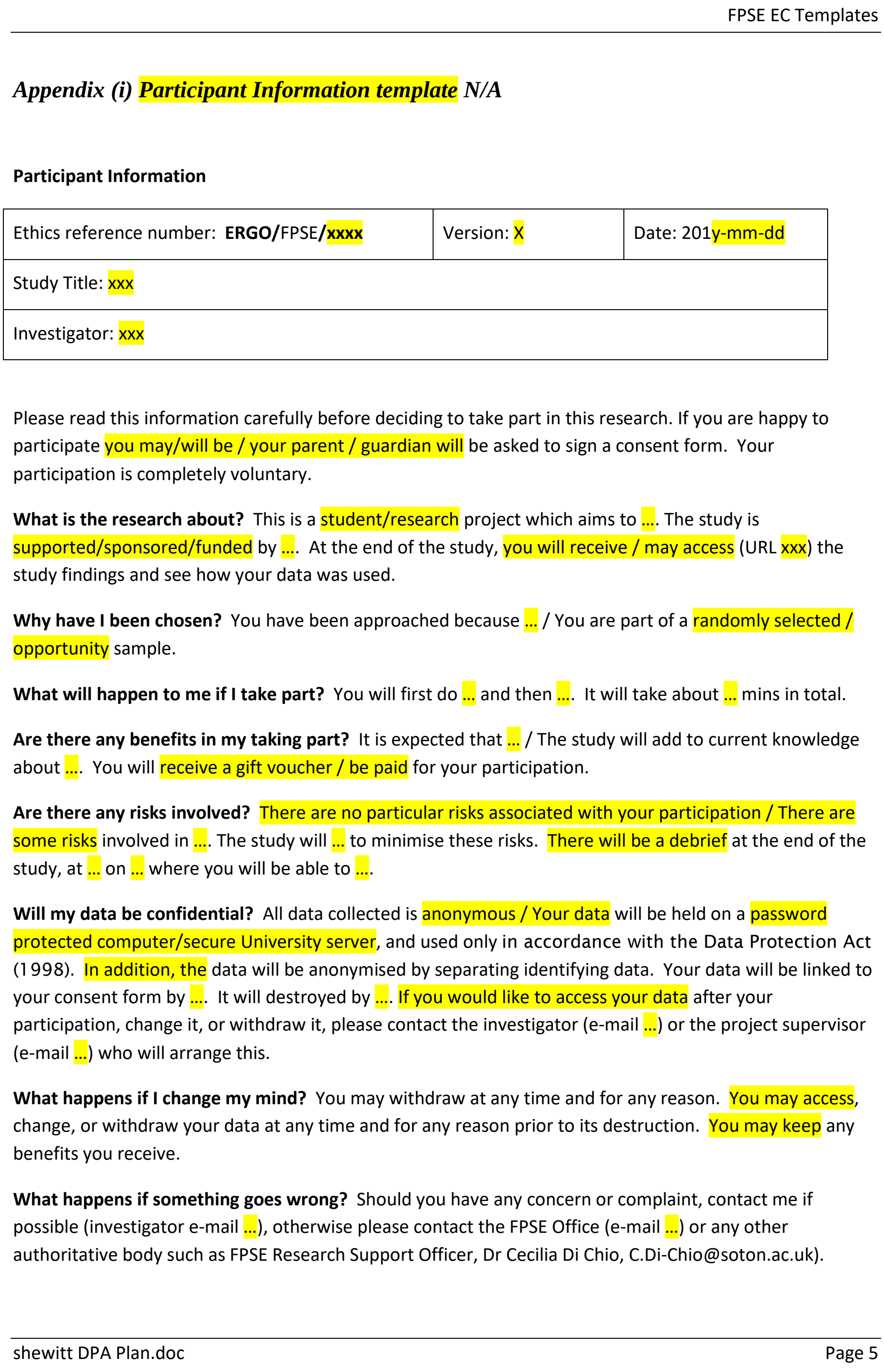
http://www.southampton.ac.uk/library/research/researchdata/

Investigators may find the University’s survey platform useful:

https://www.isurvey.soton.ac.uk/

Contacts

FPSE Research Support Officer, currently Dr Cecilia Di Chio, C.Di-Chio@soton.ac.uk.



Appendix (i) Participant Information template N/A

Participant Information

Ethics reference number: **ERGO/**FPSE/**xxxx**	Version: X	Date: 201y-mm-dd
Study Title: xxx		
Investigator: xxx		

Please read this information carefully before deciding to take part in this research. If you are happy to participate you may/will be / your parent / guardian will be asked to sign a consent form. Your participation is completely voluntary.

What is the research about? This is a student/research project which aims to …. The study is supported/sponsored/funded by …. At the end of the study, you will receive / may access (URL xxx) the study findings and see how your data was used.

Why have I been chosen? You have been approached because … / You are part of a randomly selected / opportunity sample.

What will happen to me if I take part? You will first do … and then …. It will take about … mins in total.

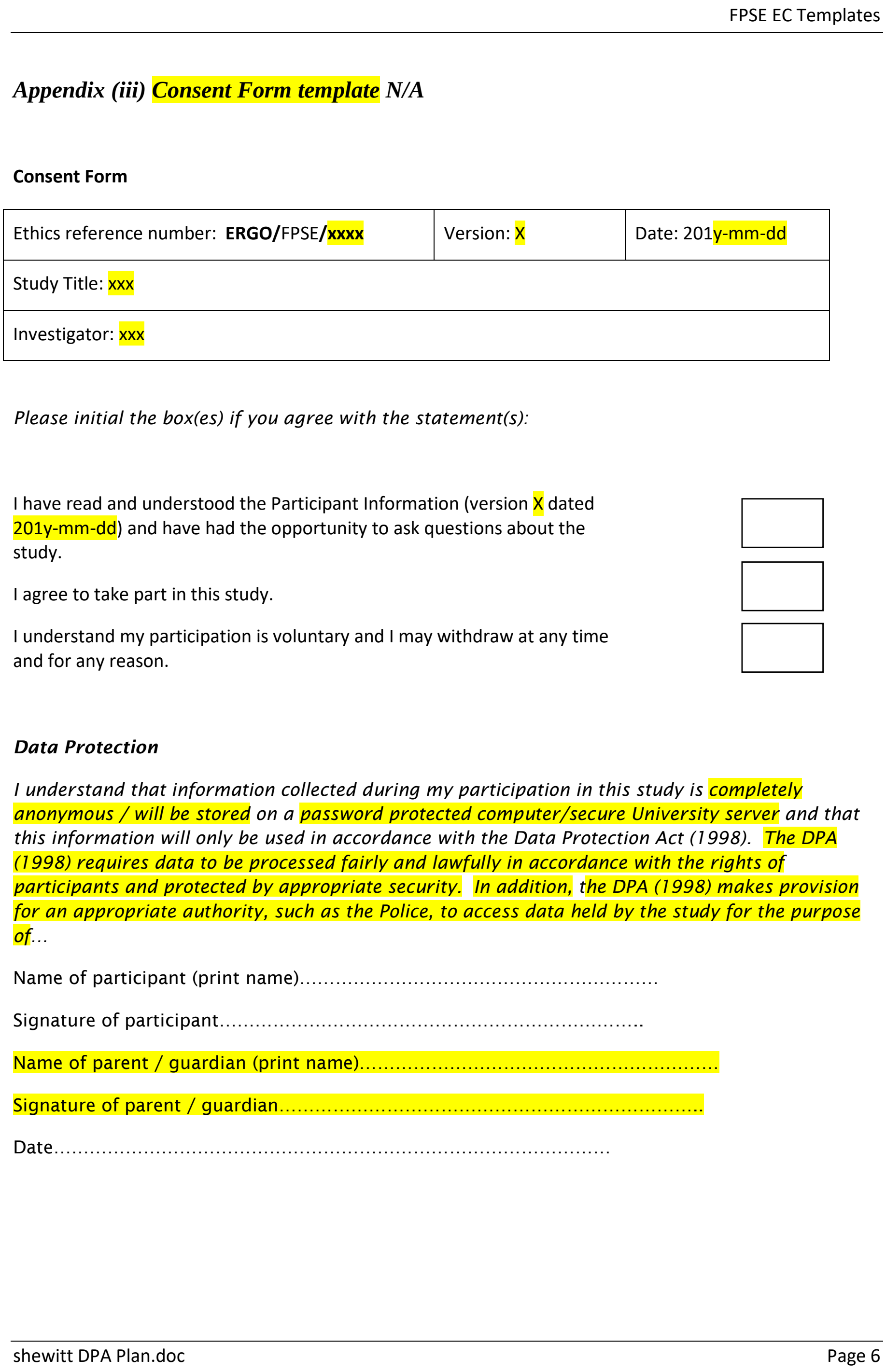
Are there any benefits in my taking part? It is expected that … / The study will add to current knowledge about …. You will receive a gift voucher / be paid for your participation.

Are there any risks involved? There are no particular risks associated with your participation / There are some risks involved in …. The study will … to minimise these risks. There will be a debrief at the end of the study, at … on … where you will be able to ….

Will my data be confidential? All data collected is anonymous / Your data will be held on a password protected computer/secure University server, and used only in accordance with the Data Protection Act (1998). In addition, the data will be anonymised by separating identifying data. Your data will be linked to your consent form by …. It will destroyed by …. If you would like to access your data after your participation, change it, or withdraw it, please contact the investigator (e-mail …) or the project supervisor (e-mail …) who will arrange this.

What happens if I change my mind? You may withdraw at any time and for any reason. You may access, change, or withdraw your data at any time and for any reason prior to its destruction. You may keep any benefits you receive.

What happens if something goes wrong? Should you have any concern or complaint, contact me if possible (investigator e-mail …), otherwise please contact the FPSE Office (e-mail …) or any other authoritative body such as FPSE Research Support Officer, Dr Cecilia Di Chio, C.Di-Chio@soton.ac.uk).



Appendix (iii) Consent Form template N/A

Consent Form

Ethics reference number: **ERGO/**FPSE/**xxxx**	Version: X	Date: 201y-mm-dd
Study Title: xxx		
Investigator: xxx		

Please initial the box(es) if you agree with the statement(s):

I have read and understood the Participant Information (version X dated 201y-mm-dd) and have had the opportunity to ask questions about the study. □

I agree to take part in this study. □

I understand my participation is voluntary and I may withdraw at any time and for any reason. □

Data Protection

I understand that information collected during my participation in this study is completely anonymous / will be stored on a password protected computer/secure University server and that this information will only be used in accordance with the Data Protection Act (1998). The DPA (1998) requires data to be processed fairly and lawfully in accordance with the rights of participants and protected by appropriate security. In addition, the DPA (1998) makes provision for an appropriate authority, such as the Police, to access data held by the study for the purpose of…

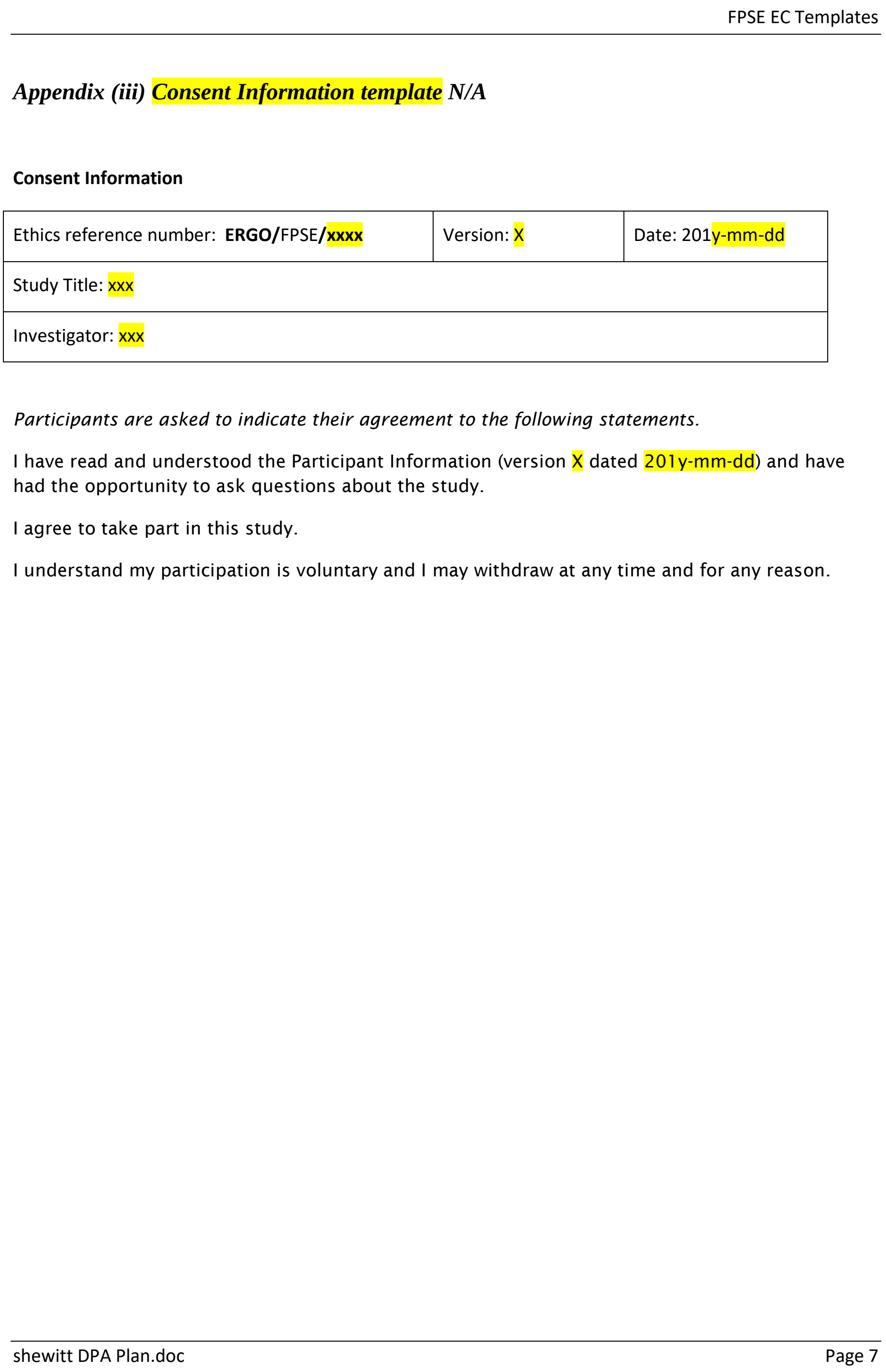
Name of participant (print name)……………………………………………………………

Signature of participant…………………………………………………………………

Name of parent / guardian (print name)…………………………………………………………

Signature of parent / guardian……………………………………………………………………

Date………………………………………………………………………………………



FPSE EC Templates

Appendix (iii) Consent Information template N/A

Consent Information

Ethics reference number: **ERGO/**FPSE**/xxxx**	Version: X	Date: 201y-mm-dd
Study Title: xxx		
Investigator: xxx		

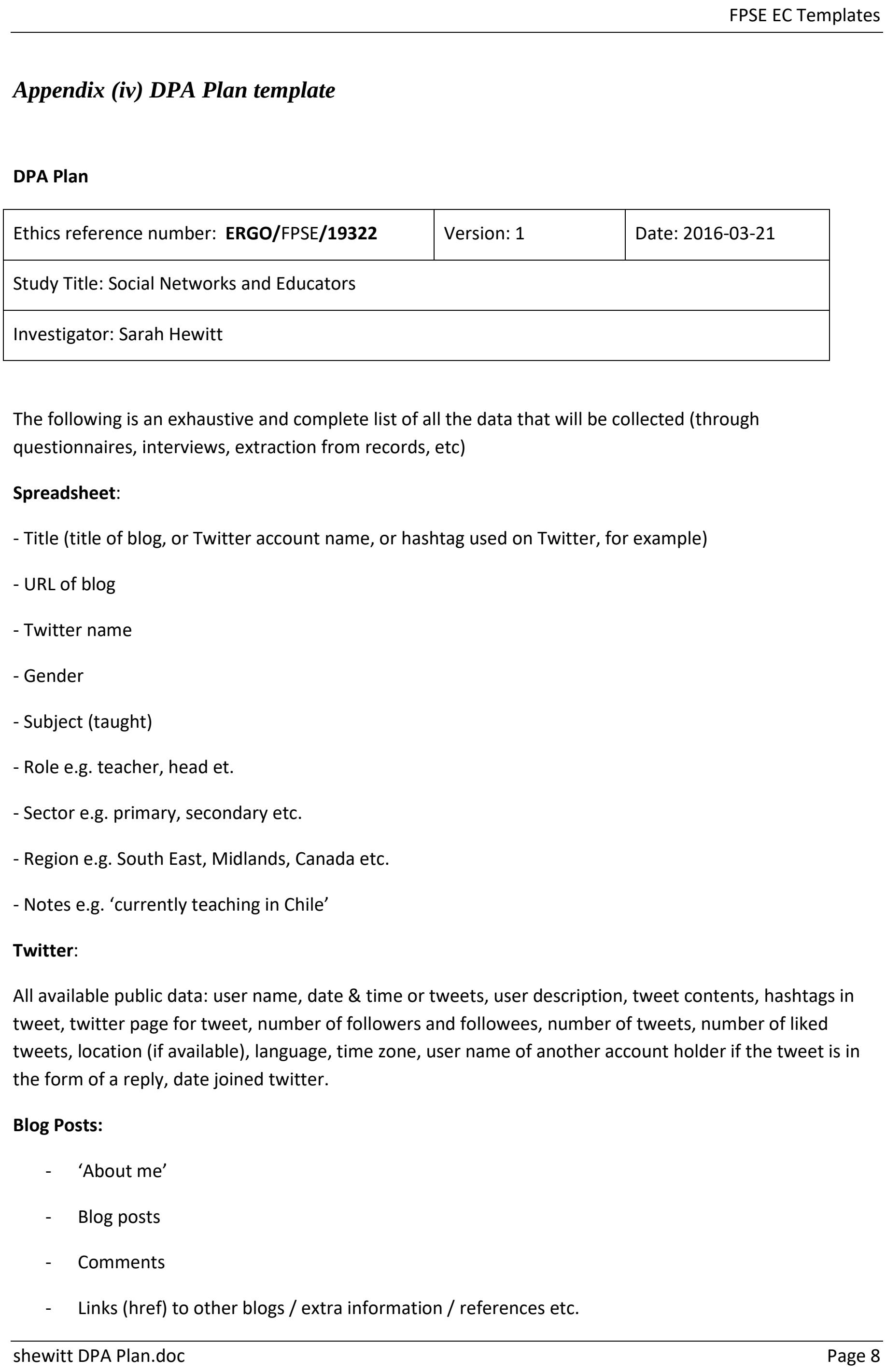
Participants are asked to indicate their agreement to the following statements.

I have read and understood the Participant Information (version X dated 201y-mm-dd) and have had the opportunity to ask questions about the study.

I agree to take part in this study.

I understand my participation is voluntary and I may withdraw at any time and for any reason.

shewitt DPA Plan.doc Page 7



Appendix (iv) DPA Plan template

DPA Plan

Ethics reference number: **ERGO/**FPSE**/19322**	Version: 1	Date: 2016-03-21
Study Title: Social Networks and Educators		
Investigator: Sarah Hewitt		

The following is an exhaustive and complete list of all the data that will be collected (through questionnaires, interviews, extraction from records, etc)

Spreadsheet:

- Title (title of blog, or Twitter account name, or hashtag used on Twitter, for example)

- URL of blog

- Twitter name

- Gender

- Subject (taught)

- Role e.g. teacher, head et.

- Sector e.g. primary, secondary etc.

- Region e.g. South East, Midlands, Canada etc.

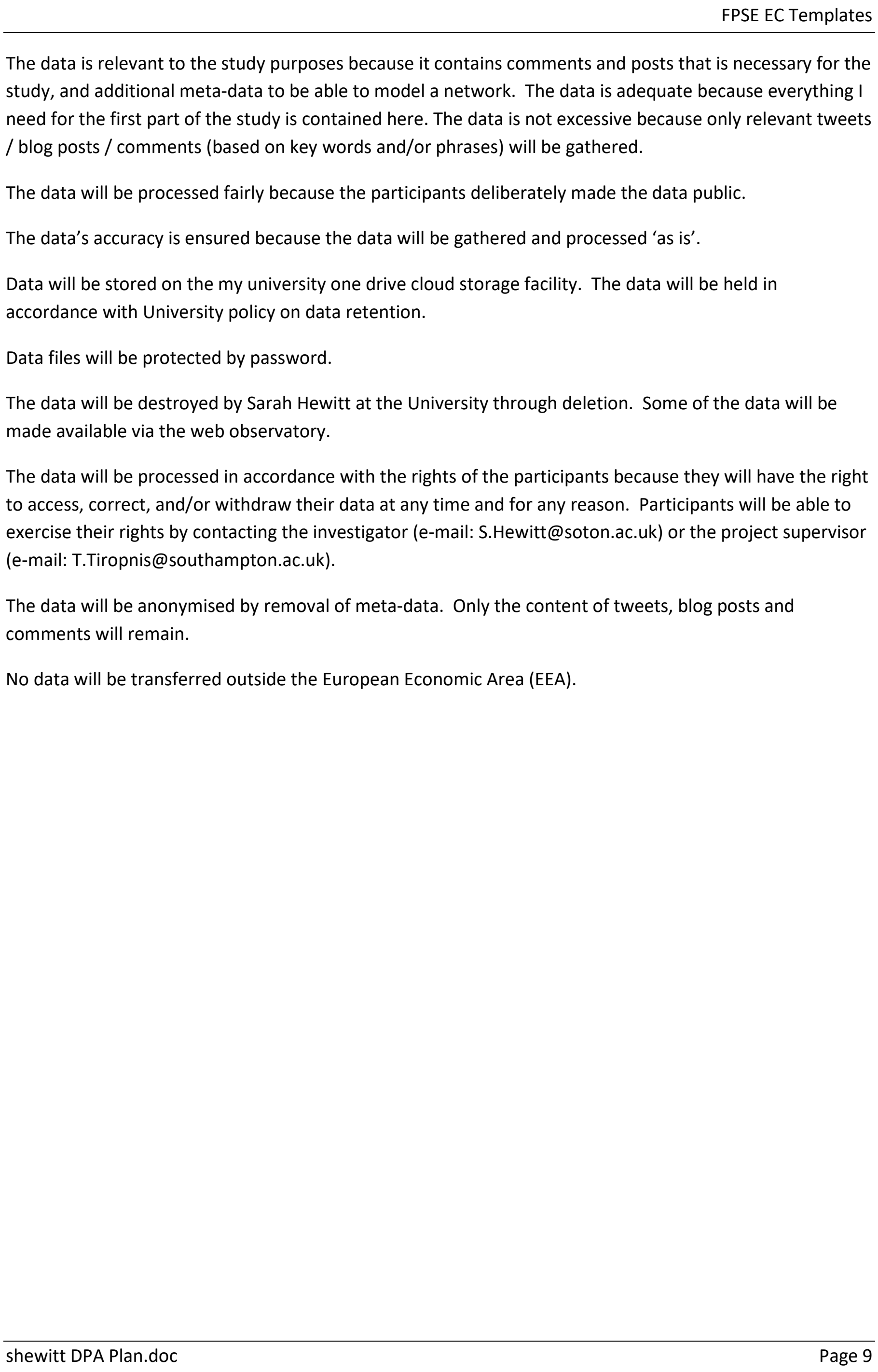
- Notes e.g. ‘currently teaching in Chile’

Twitter:

All available public data: user name, date & time or tweets, user description, tweet contents, hashtags in tweet, twitter page for tweet, number of followers and followees, number of tweets, number of liked tweets, location (if available), language, time zone, user name of another account holder if the tweet is in the form of a reply, date joined twitter.

Blog Posts:

- ‘About me’
- Blog posts
- Comments
- Links (href) to other blogs / extra information / references etc.



The data is relevant to the study purposes because it contains comments and posts that is necessary for the study, and additional meta-data to be able to model a network. The data is adequate because everything I need for the first part of the study is contained here. The data is not excessive because only relevant tweets / blog posts / comments (based on key words and/or phrases) will be gathered.

The data will be processed fairly because the participants deliberately made the data public.

The data's accuracy is ensured because the data will be gathered and processed 'as is'.

Data will be stored on the my university one drive cloud storage facility. The data will be held in accordance with University policy on data retention.

Data files will be protected by password.

The data will be destroyed by Sarah Hewitt at the University through deletion. Some of the data will be made available via the web observatory.

The data will be processed in accordance with the rights of the participants because they will have the right to access, correct, and/or withdraw their data at any time and for any reason. Participants will be able to exercise their rights by contacting the investigator (e-mail: S.Hewitt@soton.ac.uk) or the project supervisor (e-mail: T.Tiropnis@southampton.ac.uk).

The data will be anonymised by removal of meta-data. Only the content of tweets, blog posts and comments will remain.

No data will be transferred outside the European Economic Area (EEA).

References

Aggarwal, C. C. and Zhai, C. (2012). A survey of text clustering algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining text data*, pages 78 – 128. Springer Science+Business Media, United States.

Agrawal, A., Fu, W., and Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, 98:74–88.

Bakharia, A. (2019). On the Equivalence of Inductive Content Analysis and Topic Modeling. In *Communications in Computer and Information Science*, volume 1112, pages 291–298. Springer.

Balakrishnan, V. and Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, 2(3):262–267.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., and Weitzner, D. J. (2006). A Framework for Web Science. *Foundations and Trends in Web Science*, 1(1):1–130.

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information*, 29(4):308–316.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blood, R. (2004). How blogging software reshapes the online community. *Communications of the ACM*, 47(12):53.

Booth, S. E. and Kellogg, S. B. (2014). Value creation in online communities for educators. *British Journal of Educational Technology*, 46(4):684–698.

Carpenter, J. P. and Krutka, D. G. (2014). How and Why Educators Use Twitter: A Survey of the Field. *Journal of Research on Technology in Education*, 464:1539–1523.

Carpenter, J. P. and Krutka, D. G. (2017). Professional Development in Education Engagement through microblogging : educator professional development via Twitter. *Professional Development in Education*, 41(4):707–728.

Cho, V., Ro, J., and Littenberg-Tobias, J. (2013). What Twitter will and will not do: theorizing about teachers' online professional communities. *LEARNing Landscapes —*, 6(2).

Cohen, L., Manion, L., and Morrison, K. (2007). *Research Medhods In Education.* Routledge, Taylor & Francis, sixth edition.

Creswell, J. W. and Clark, V. L. P. (2018). *Designing and conducting mixed methods research.* Los Angeles ; London : SAGE, 2018, third edition.

Davis, K. (2015). Teachers' perceptions of Twitter for professional development. *Disability and Rehabilitation*, 37(17):1551–1558.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6):570–606.

Duncan-Howell, J. (2010). Teachers making connections: Online communities as a source of professional learning. *British Journal of Educational Technology*, 41(2):324–340.

Eagan, B., Misfeldt, M., and Siebert-Evenstone, A., editors (2019). *Advances in Quantitative Ethnography.* Springer International Publishing.

Eickhoff, M. and Wieneke, R. (2018). Understanding Topic Models in Context: A Mixed-Methods Approach to the Meaningful Analysis of Large Document Collections. In *51st Hawaii International Conference on System Sciences*, pages 903–912.

Foster, D. (2018). Teacher recruitment and retention in England. *House of Commons Library*, (7222).

Galyardt, A., Aleahmad, T., Fienberg, S., Junker, B., and Hargadon, S. (2009). Analysis of a Web-based Network of Educators. *Carnegie Mellon University*, pages 1–31.

Giddens, A. (1984). *The Constitution of Society Outline of the Theory of Structuration.* University of California Press (Berkeley and Los Angeles).

Gillard, D. (2015). Gove v . the Blob : the Coalition and education. *Forum*, 57(3):277–294.

Greene, K. (2013a). Investigating Teacher Voice Through Blogs: Policy, Practice, and Local Knowledge. *Selected Papers of Internet Research.*

Greene, K. (2013b). *NOTES FROM THE BLOGGING FIELD: TEACHER VOICE AND THE POLICY-PRACTICE GAP IN EDUCATION.* PhD thesis.

Greene, K. (2016). Teacher blogs and education policy in a publicly private world: filling the gap between policy and practice. *Learning, Media and Technology.*

Hallsworth, M., Parker, S., and Rutter, J. (2011). Policy Making in the Real World. *Institute for Government*, pages 1–105.

Hopkins, D. J. and King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, 54(1):229–247.

Hou, H. T., Chang, K. E., and Sung, Y. T. (2010). What kinds of knowledge do teachers share on blogs? A quantitative content analysis of teachers' knowledge sharing on blogs. *British Journal of Educational Technology*, 41(6):963–967.

Hou, H. T., Chang, K. E., and Sung, Y. T. (2011). A longitudinal analysis of the behavioural patterns in teachers using blogs for knowledge interactions. *British Journal of Educational Technology*, 42(2):2010–2012.

Hur, J. W. and Brush, T. A. (2009). Teacher Participation in Online Communities: Why Do Teachers Want to Participate in Self-generated Online Communities of K–12 Teachers? *Journal of Research on Technology in Education JRTE*, 279336(413):279–303.

Jacobi, C., Van Atteveldt, W., and Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalish.*

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.

Johnson, C. (2001). A survey of current research on online communities of practice. *The Internet and Higher Education*, 4(1):45–60.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68.

Kelly, N. and Antonio, A. (2016). Teacher peer support in social network sites. *Teaching and Teacher Education*, 56:138–149.

Kidd, W. (2013). Investigating the lives of new teachers through ethnographic blogs. *Ethnography and Education*, 8(2):210–223.

Kirby, D. and Cameron, M. (2008). Blogging in Academe. *Academic Matters*, (January 2008):16–20.

Kirkup, G. (2010). Academic blogging: academic practice and academic identity. *London Review of Education*, 8(1):75–84.

Krippendorff, K. (2004). *Content Analysis. An Introduction to its Methodology.* Sage Publications Inc., 2nd edition.

Krutka, D. G., Carpenter, J. P., and Trust, T. (2017). Enriching Professional Learning Networks: A Framework for Identification, Reflection, and Intention. *TechTrends*, 61(3):246–252.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35.

Lantz-Andersson, A., Lundin, M., and Selwyn, N. (2018). Twenty years of online teacher communities: A systematic review of formally-organized and informally-developed professional learning groups. *Teaching and Teacher Education*, 75:302–315.

Lave, J. and Wenger, E. (1991). *Situated learning : legitimate peripheral participation.* Cambridge University Press.

Lee, S. B., Gui, X., Manquen, M., and Hamilton, E. R. (2019). Use of Training, Validation, and Test Sets for Developing Automated Classifiers in Quantitative Ethnography. In *Communications in Computer and Information Science*, volume 1112, pages 117–127. Springer.

Liu, H., Christiansen, T., Baumgartner, W. A., Verspoor, K., and Verspoor, K. (2012). BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*, 3:3.

Lo, R. T.-w., He, B., and Ounis, I. (2005). Automatically Building a Stopword List for an Information Retrieval System. *JDIM*, 3:2–8.

Lough, C. (2019). GCSE results 2019: Schools battling for 'forgotten third' — Tes.

Loving, C. C., Schroeder, C., Kang, R., Shimek, C., and Herbert, B. (2007). Blogs: Enhancing links in a professional learning community of science and mathematics teachers. *Contemporary Issues in Technology and Teacher Education*, 7(3):178–198.

Luehmann, A. L. (2008a). Blogs' Affordances for Identity Work: Insights Gained From an Urban Teacher's Blog. *The New Educator*, 4(3):175–198.

Luehmann, A. L. (2008b). Using Blogging in Support of Teacher Professional Identity Development: A Case Study. *Journal of the Learning Sciences*, 17(3):287–337.

Luhn, H. P. (1960). Keyword-in-Context Index for Technical Literature (KWIC). *American Documentation*, 11(4):288–295.

Macia, M. and Garcia, I. (2016). Informal online communities and networks as a source of teacher professional development: A review. *Teaching and Teacher Education*, 55:291–307.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Mikhaylov, S., Laver, M., and Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91.

Moreno-Sánchez, I., Font-Clos, F., and Corral, Á. (2016). Large-scale analysis of Zipf's law in English texts. *PLoS ONE*, 11(1):5–7.

Murakami, A., Thompson, P., Hunston, S., and Vajn, D. (2017). 'What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*, 12(2):243–277.

Nardi, B., Schiano, D., and Gumbrecht, M. (2004). T 18: Blogging as social activity, or, would you let 900 million people read your diary? *... of the 2004 ACM conference on ...*, pages 222–231.

Nothman, J., Hanmin, Q., and Yurchak, R. (2018). Stop Word Lists in Free Open-source Software Packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12. Association for Computational Linguistics.

Peal, R. (2014). Playing the Game The enduring influence of the preferred Ofsted teaching style. (July).

Peal, R., editor (2015). *Changing Schools. Perspective on five years of educational reform.* John Catt Educational Ltd, Woodbridge, UK, first edition.

Phethean, C., Simperl, E., Tiropanis, T., Tinati, R., and Hall, W. (2016). The Role of Data Science in Web Science. *IEEE Intelligent Systems*, 31(3):102–107.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future Directions. *National Institutes of Health*, 21(October):1112 – 1130.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, pages 130–137.

Porter, M. (2001). Snowball: A language for stemming algorithmsl.

Ray, B. B. and Hocutt, M. M. (2006). Teacher-centered Weblogs: Perceptions and Practices. *Journal of Computing in Teacher Education*, 23(1):11–18.

Rheingold, H. (1993). The virtual community. *http://www.rheingold.com/vc/book/intro.html.*

Robson, J. (2016). Engagement in structured social space: an investigation of teachers' online peer-to-peer interaction. *Learning, Media and Technology*, 41(4):119–129.

Robson, J. (2017). Performance, structure and ideal identity: Reconceptualising teachers' engagement in online social spaces. *British Journal of Educational Technology*, 00(00).

Russell, M. A. (2014). *Mining The Social Web.* O'Reilly Media Inc., second edition.

Salganik, M. J. (2018). Bit By Bit: Social Research in the Digital Age.

Schaffer, W. (2017). *Quantitative Ethnography.* Cathcart Press, Madison, Wisconsin.

Schau, H.J. & Gilly, M.C. (2003). We are what we post? Self- presentation in personal web space. *Journal of Consumer Research*, 30(3):385–404.

Schiano, D. J., Nardi, B. a., Gumbrecht, M., and Swartz, L. (2004). Blogging by the rest of us. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1143–1146.

Schmidt, J. (2007). *Blogging practices: An analytical framework*, volume 12. Wiley-Blackwell.

Scholar, T. (2018). The Green Book.

Sievert, C. and Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70. Association for Computational Linguistics.

Silva, C. and Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, 3:1661–1666.

Symonds, J. E. and Gorard, S. (2009). The Death of Mixed Methods: Research Labels and their Casualties. In *The British Educational Research Association Annual Conference*, Edinburgh.

Tashakkori, A. and Teddlie, C. (2010). *Sage handbook of Mixed Methods in Social & Behavioral Research.* Sage Publications Inc., London, 2nd edition.

Tour, E. (2017). Teachers' self-initiated professional learning through Personal Learning Networks. *Teaching, Pedagogy and Education*, 26(2):179–192.

Trant, J. (2009). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, 10(1).

Trust, T., Krutka, D. G., and Carpenter, J. P. (2016). "Together we are better": Professional learning networks for teachers. *Computers & Education*, 102:15–34.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Van Rijsbergen, C. (1979). *Information Retrieval.*

Visser, R. D., Calvert Evering, L., and Barrett, D. E. (2014). #TwitterforTeachers: The Implications of Twitter as a Self-Directed Professional Development Tool for K–12 Teachers. *Journal of Research on Technology in Education*, 464:1539–1523.

Wellman, B. (2001). Physical place and cyberplace: The rise of personalized networking. *International Journal of Urban and Regional Research*, 25(2):227–252.

Wellman, B. and Gulia, M. (1999). Virtual communities as communities. *Communities in Cyberspace*, pages 167–193.

Wesely, P. M. (2013). Investigating the Community of Practice of World Language Educators on Twitter. *Journal of Teacher Education*, 64(4):305–318.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study of Feature selection in Text Categorisation. *Bioinformatics*.

Zhu, X. and B, G. A. (2009). *Introduction to Semi-Supervised Learning*. Morgan & Claypool, first edition.

Zipf, G. K. (1966). *Psychobiology of Language*. The M.I.T. Press.