The impact of sentiment on school inspection reports in England

Objectives or purposes

For several decades already, the national inspectorate for England (Ofsted) aims to improve schools by inspecting them and offering a diagnosis of what they should do to improve. Implicit in this aim is the assumption that Ofsted can reliably judge the quality of education being offered by a school. To this end, schools are periodically judged. While Ofsted do state criteria against which schools are judged, they do not explicitly state the *practices* which they consider to be associated with high (or low) standards of education, meaning inspectors have to exercise discretion when reaching a judgement. Given that Ofsted employs thousands of inspectors, this raises the question of whether inspectors look for the same thing when judging a school, or indeed offer consistent advice. We analyse the complete corpus of nationally published Ofsted inspection reports since 2000 to empirically test the research question whether inspection report sentiments have changed over time with the coming and going of different Head Inspectors. We utilise text mining and sentiment analysis to explore these questions. The scholarly significance of the study is that we gain insight into the inspection process, which can contribute to reducing the ambiguity in the inspection process.

Perspective(s) or theoretical framework

The Office for Standards in Education (Ofsted) was created in 1992. Her Majesty's Inspectors continued to be crown-appointees but were now located in Ofsted, an independent, non-ministerial government department. The new inspectorate explicitly aimed for "improvement through inspection" by holding schools and local authorities to account (Marshall, 2008). The information necessary to hold schools to account in this way was collected through inspecting all schools once every four years using a team of approximately fifteen inspectors, including subject specialists, who spent four or five days at the school collecting evidence (Elliott, 2012). As well as adopting the objective of improving schools, Ofsted also adopted an explicit framework for judging schools which states explicit criteria for what constituted a good school, with all schools graded from 1 (excellent) to 7 (very poor). Despite this explicit, common framework, Ofsted recognised that inspectors judgement still played an substantial role, with the inspection handbook noting that "The basic principle has always been close observation exercised with an open mind by persons with appropriate experience and a framework of relevant principles." (Matthews & Sammons, 2004, p.82). Concerns about the consistency of the inspection process began to arise soon after Ofsted was established (Maw, 1995; Sinkinson & Jones, 2001; Penn, 2002). One reason for this was that, in order to inspect all schools once every four years, Ofsted had to recruit 7,500 contracted inspectors in the two years after it was established (Elliott, 2012). New concerns emerged in 2005 when the inspection process underwent radical reform. The size of the inspection teams were cut from around fifteen to just three or four, which meant they could no longer accommodate subject specialists. The length of inspections were also cut from five down to two days each. Inspections were therefore reduced from approximately seventy-five man-days, to approximately eight. This made the model of inspection based on "close observation exercised with an open mind by persons with appropriate experience" infeasible. The responsibility for collecting evidence therefore shifted from inspectors to schools (Elliott, 2012). Another consequences of these reforms was a greater reliance on schools self-evaluating and creating an auditable paper trail of evidence ready for inspectors to review when they arrived (Plowright, 2007; Allen & Sims, 2018). In 2010, the Conservatives came into government and a number of changes were made to the way Ofsted operates, characterised by Elliott (2012) as "slimming down and toughening up". Slimming down involved reducing the number of inspection criteria from 29 to just 4, ending the practice of grading teaching in individual lessons, and increasing the number of inspections conducted by serving school leaders, rather than specialist consultants (Baxter & Clarke, 2013; Cladingbowl, 2014). Toughening up refers to the policy of forcing all schools that are rated 'Inadequate' by Ofsted to convert to academy status, which generally involves the headteacher losing their job (Eyles & Machin, 2015). The reforms therefore reduced clarity about how Ofsted judged schools at the same time as raising the stakes of such judgments. It is perhaps unsurprising that speculation and rumours began to emerge about what Ofsted required from schools. Ofsted were criticised for allowing this ambiguity to persist, contributing to high levels of teacher workload in the process, as cautious headteachers insisted on certain planning and marking practices which Ofsted were rumoured to favour, in order to minimise the risk of being given a bad Ofsted grade (TNS BRMB, 2014; Allen & Sims, 2018).

Given the suspected ambiguity in inspection judgements, we wanted to explore whether, since 2000, inspection report sentiments have changed over time; might any changes might be related to the Head Inspector at the time, as displayed in Table 1.

Table 1 Head Inspectors since 2000 to now.

HMCI	In office	Grouping	#
Mike Tomlinson	2000-2002	2000-2002	712
Sir David Bell	2002–2006	2003-2006	1492
Maurice Smith	January 2006–October 2006 (acting)		
Christine Gilbert	2006–2011	2007-2011	5220
Miriam Rosen	July 2011–December 2011 (acting)		
Sir Michael Wilshaw	January 2012–December 2016	2012-2016	8881
Amanda Spielman	January 2017–present	2017	907
		Total	17,212

The research question at hand, then is: Do England's inspection reports by Ofsted show changes in sentiment over time? By studying this, we get insight into the inspection process, which can be used to perhaps reduce the ambiguity in the inspection process.

Methodology

We want to study inspection reports since 2000, and this entails analysing hundreds of thousands of words at scale, making data mining, and text mining in particular an appropriate method. To structure the process of 'knowledge discovery in data' we utilise a standardised procedure for data mining, the Cross Industry standard Process for Data Mining (CRISP-DM). This procedure distinguishes several phases that could be applied to the resulting big data. The first phase, Organizational Understanding, concerns an understanding of the web data: what data is actually on the web, what does it say, and how could it be useful for us. The second phase, Data Understanding, would involve knowing the precise format of the data. In phase three, Data Preparation, the data is transformed into a format that is understandable for the tools that will perform the analyses. Phase four, Modelling, is the phase that is used for the actual analyses. Phase five, Evaluation, determines the truthfulness and usefulness of the analysis results by providing some interpretation of the model results. Finally, phase six, Deployment, could involve the distribution and publication of the results of the analyses, as is done in this paper, and therefore not explicitly mentioned. We now describe the steps undertaken in every phase. As the phases encompass both methodological choices and results, the following sections can be seen as combined methodological and results sections.

Phase 1: Organizational understanding

Ofsted provides publicly-available inspection reports and other inspection documents for every school on their website. These are available from 2000. There are different types of reports, ranging from full inspection reports to shorter interim documents. We decided to include all the documents, as they all say something about the way the inspection operates. This specific analysis, unlike a prior analysis in Bokhove (2015), does not utilise the specific judgements, but takes into account all text produced by Ofsted.

Phase 2: Data collection and data understanding

A web scraper was set up with the browser extension Web Scraper (http://webscraper.io/) and used to scrape the Ofsted website at http://www.ofsted.gov.uk/. The scraper collected the URLs of all historical inspection reports and other reports since the year of first publication, 2000 (N=17,212, 2.49 GB of data). A mass downloader was subsequently used to download all the PDF documents. A complete overview of the scrape is presented in Table 2. The scrape was performed towards the end of 2017, excluding the last months of that last year. The dataset allows us to both look at the most current inspection reports, and any differential effects, as well as changes over time. All files were in PDF format.

Table 2 Total number and size of inspection documents from 2000 to 2017

Year	Number of documents	Size in Mb
2000	228	36.9
2001	287	40.4
2002	197	33.8
2003	212	37.8
2004	303	39.1
2005	310	51.7
2006	667	60.3
2007	961	133.0
2008	888	140.0
2009	939	135.0
2010	1113	159.0
2011	1319	223.0
2012	2521	324.0
2013	2125	331.0
2014	1740	274.0
2015	1419	225.0
2016	1076	170.0
2017	907	138.0
Total	17212	2.49 Gb

Phase 3: Data preparation

In this phase the data were prepared for sentiment analysis. To do this we imported all the PDF files into Rstudio, a free and open-source integrated development environment for R, a programming language for statistical computing and graphics (www.rstudio.com). PDFs were grouped by period of head inspector as indicated in Table 1. Within Rstudio we converted all the PDF documents to a so-called 'tidy text format', which consists of a table with one-token-per-row (Silge & Robinson, 2017). For this we used the tidytext package in Rstudio (https://www.tidytextmining.com/).
Practically, this meant that all reports were broken up in separate words, with every word part of a report within one of the Table 1 groupings. This resulted in a table with 32,235,414 rows (one per word) as a basis for further analysis.

Phase 4: Modelling

We want to explore the sentiments in the processed inspection documents. Human readers would use understanding of the emotional intent of words to conclude whether a section of text is positive or negative, or in other cases perhaps more nuance emotions like surprise and disgust. In text mining the emotional content of a text can be explored algorithmically. Within the tidy text realm this workflow is depicted in Figure 1.

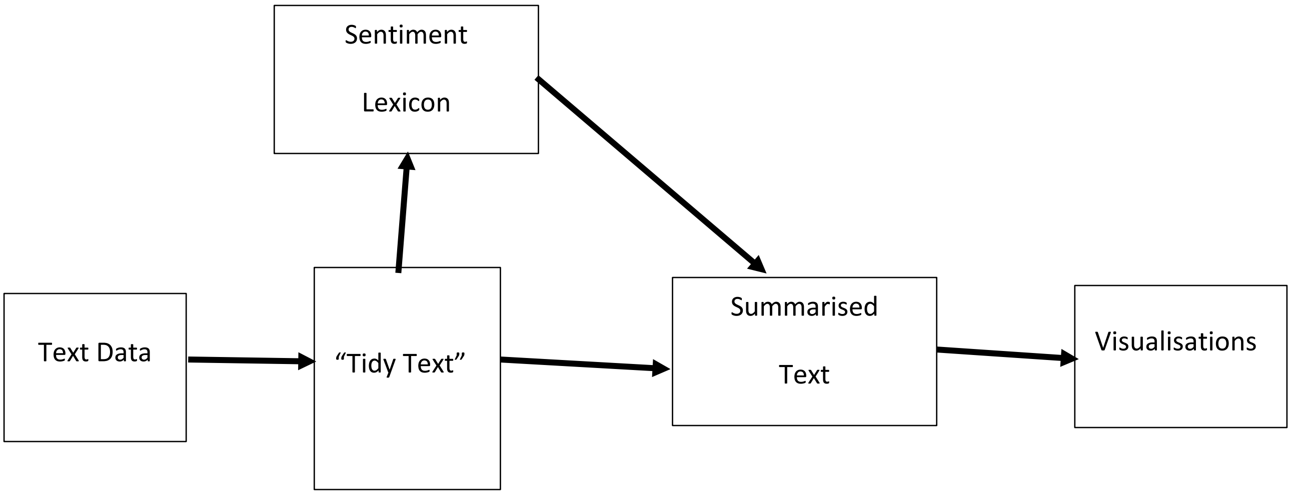
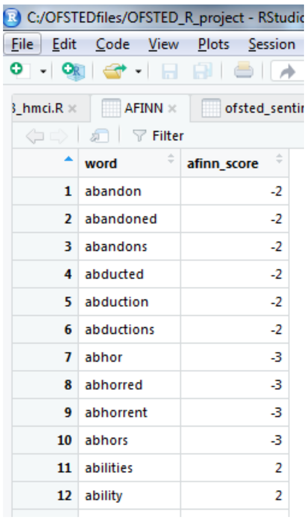


Figure 1 Depiction of the flowchart of a typical text analysis that uses tidytext for sentiment analysis. Adapted from Silge and Robinson (2017).

A common way to analyse the sentiment of a text is to break down the text as a combination of its individual words and the sentiment content of a text and even a corpus of texts, as the sum of the sentiment content of the individual words. To apply sentiment analysis, as depicted as well in Figure 1, a sentiment lexicon is used. The general-purpose AFINN lexicon (Nielsen, 2011) was utilised. The lexicon is based on unigrams, single words, and contains English words and the words are assigned scores for positive/negative sentiment, and also possibly emotions like joy, anger, sadness, etcetera. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. Figure 2 shows a fragment of the lexicon.


C:/OFSTEDfiles/OFSTED_R_project - RStudic
File Edit Code View Plots Session
3_hmci.R
AFINN
ofsted_sentir
Filter

	word	afinn_score
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2

Figure 2 Fragment of the AFINN lexicon in R studio.

These scores allow us to summarise our texts or a corpus of texts, which then can be visualised, as done in the next phase. The documents were grouped according to year, and according to 'Head Inspector' phases as described in Table 1.

Phase 5: Evaluation

In this phase we present some of the results obtained and provide some interpretation. Figure 2 shows that how sentiments by Head Inspector over the period 2000 to 2017 have changed, starting just under an average sentiment score of 1.3 and rising to over 1.6 under Head Inspector Christine Gilbert. There is a large overlap with the Labour government. After that, a Conservative government came into play, the average sentiment score decreased again, with the last Head Inspector scoring even under the 2000-2002 score.

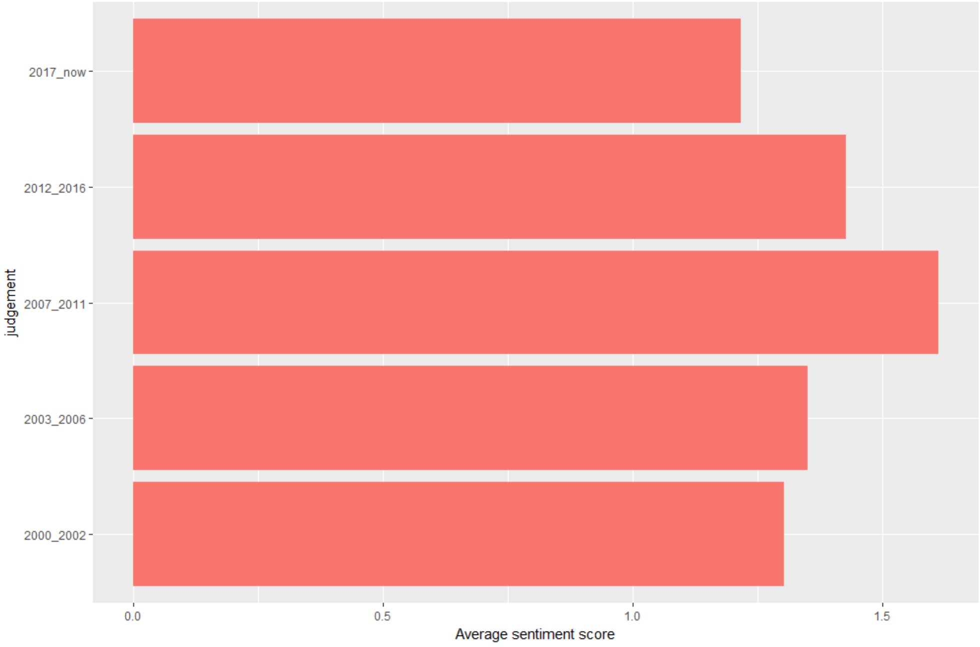


Figure 3 Average sentiment score for the corpus of inspection documents by Head Inspector.

A more fine-grained analysis of the words that contributed to the sentiment scores is presented in Figure 4. Notable observations are that certain words contributed to sentiment scores throughout 2000 to 2012, for example 'progress', 'improvement' and 'support'. There also, however, are differences, for example the use of the word 'care', and also the negative influence of the word 'disadvantaged' in the last inspection year.

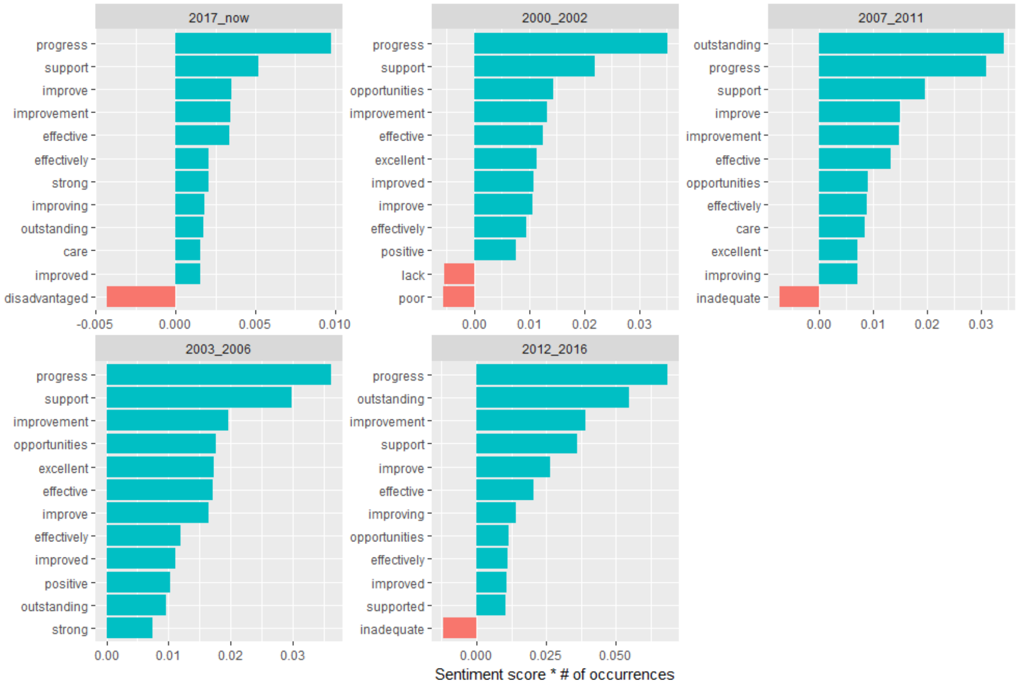


Figure 4 Twelve words that contributed the most to sentiment scores within each corpus of inspection documents by Head Inspector.

A more detailed analysis on a year-by-year basis confirmed the trajectory of sentiments starting lower in 2000 then rising slowly to its highest point in 2011, after which it decreased again swiftly towards 2017, the lowest average in the complete corpus of 17,212 inspection documents. The trajectory of sentiments coincides with changes in the inspection regime, that aimed to make the process more challenging, especially for ‘coasting schools’, schools that underperform (Department for Education, 2012). In that respect we can tentatively see a relationship between the sentiments expressed in the inspection reports and the policy level decisions as expressed in government policy documents.

Conclusion

Sentiment analysis indeed seems to be able to describe policy changes over time. An important element to keep in mind is that interpretation of results from techniques like these are inevitably contextual by nature. Without knowing enough about the English inspection system, as well as some of the history behind it, interpretation of sentiment scores and word use will be extremely difficult. Ideally, analyses like this should be accompanied with other analysis methods so results can be triangulated.

References

Allen, R., & Sims, S. (2018). *The Teacher Gap*. London: Routledge.

Bokhove, C. (2015). *Text mining school inspection reports in England with R*. Working paper. Retrieved from https://eprints.soton.ac.uk/393849/1/websci_bokhove_FINAL.pdf

Baxter, J., & Clarke, J. (2013). Farewell to the tick box inspector? Ofsted and the changing regime of school inspection in England. *Oxford Review of Education*, *39*(5), 702-718.

Bosnjak, Z., Grljevic, O., & Bosnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. Applied Computational Intelligence and Informatics, 2009. SACI '09. 5th International Symposium on , vol., no., pp.509-514, 28-29 May 2009 doi:10.1109/SACI.2009.5136302

Cladingbowl, M. (2014). Why I want to try inspecting without grading teaching in each individual lesson. *Ofsted.* Retrieved from www.leicesterpp.org.uk/main/FILE/1339

Department for Education. (2012). *Ofsted scraps 'satisfactory' judgement to help improve education*. Retrieved from https://www.gov.uk/government/news/ofsted-scraps-satisfactory-judgement-to-help-improve-education

Elliott, A. (2012). Twenty years inspecting English schools–Ofsted 1992–2012. *Rise Review*, 1-4.

Eyles, A., & Machin, S. J. (2015). The introduction of academy schools to England's education. CEP Discussion Paper No. 1368. Retrieved from http://files.eric.ed.gov/fulltext/ED574319.pdf

Marshall, C. (2008). School inspection: Thirty-five years of school inspection: raising educational standards for children with additional needs? *British Journal of Special Education*, *35*(2), 69-77.

Maw, J. (1995). The Handbook for the Inspection of Schools: a critique. *Cambridge Journal of Education*, *25*(1), 75-87.

Matthews, P., & Sammons, P. (2004). *Improvement through inspection: An evaluation of the impact of Ofsted's work*. HMI 2244. Retrieved from http://dera.ioe.ac.uk/4969/

Nielsen, F.Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages* (pp. 93-98). http://arxiv.org/abs/1103.2903

Penn, H. (2002). 'Maintains a Good Pace to Lessons': inconsistencies and contextual factors affecting OFSTED inspections of nursery schools. *British Educational Research Journal*, *28*(6), 879-888.

Plowright, D. (2007). Self-evaluation and Ofsted inspection: developing an integrative model of school improvement. *Educational Management Administration & Leadership*, *35*(3), 373-393.

TNS BRMB (2014). *Teachers' workload diary survey 2013*. Department for Education.

Silge, J., & Robinson, D. (2018). *Text Mining with R - A Tidy Approach*. O'Reilly Media: Sebastopol, CA.

Sinkinson, A., & Jones, K. (2001). The validity and reliability of Ofsted judgements of the quality of secondary mathematics initial teacher education courses. *Cambridge Journal of Education*, *31*(2), 221-237.