Text mining school inspection reports in England with R

Christian Bokhove
Southampton Education School
University of Southampton
C.Bokhove@soton.ac.uk

ABSTRACT

This short paper reports on the first results of a text mining analysis of publicly-available OFSTED secondary school inspection reports for 1766 schools from 2000 to February 2014. The analysis focuses on what OFSTED has written in reports over this period, and how this relates to the judgment OFSTED has given to a specific school. It serves as a proof-of-concept of how text mining could convey some meaning from a vast amount of documents. The focus of this analysis is on the judgments that OFSTED makes in every report. The analysis was conducted by first 'scraping' the reports from the OFSTED website and then utilising sentiment analysis and topic modelling techniques in R to extract features of these documents. There appears to be link between the reports' judgment and the sentiments in the report, as well as differences in the topics observed. However, interpreting these findings from data mining alone was not straightforward.

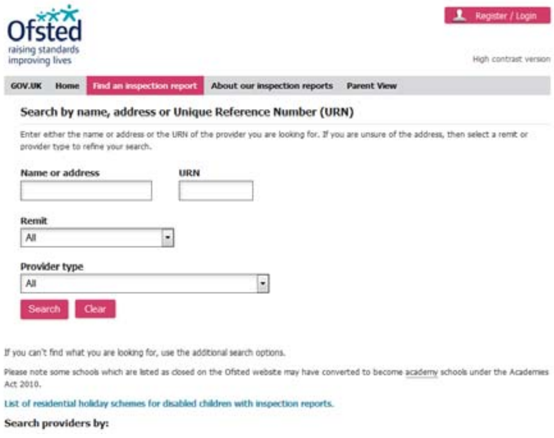
Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing – *text analysis.*

Keywords

scraping; inspection reports; OFSTED; textmining; sentiment analysis; topic modelling; lda

Figure 1: OFSTED site for inspection reports



Search by name, address or Unique Reference Number (URN)

Enter either the name or address or the URN of the provider you are looking for. If you are unsure of the address, then select a remit or provider type to refine your search.

Name or address

URN

Remit: All

Provider type: All

If you can't find what you are looking for, use the additional search options.

Please note some schools which are listed as closed on the Ofsted website may have converted to become academy schools under the Academies Act 2010.

List of residential holiday schemes for disabled children with inspection reports.

Search providers by:

1. INTRODUCTION

In England there is an important role for OFSTED, the official body for inspecting schools, when it comes to inspection of the quality of teaching. This short paper reports on the first results of a text mining analysis of the most recent publicly-available OFSTED secondary school inspection reports for 1766 schools. The reports can be found on the website http://reports.ofsted.gov.uk (Figure 1). The analysis focuses on what OFSTED has written in reports over this period, and how this relates to the judgment OFSTED has given to a specific school. It serves as a proof-of-concept of how text mining could convey some meaning from a vast amount of documents. The analysis was conducted by first 'scraping' the reports from the OFSTED website and then utilising sentiment analysis and topic modelling techniques to extract features of these documents. One interest was whether –as could be expected- there were different sentiments for the different judgments i.e. that an outstanding school would have more positive word use in their inspection report than a school 'requiring improvement'. As the name of the judgments have changed slightly over the years it also is interesting to see how this might relate with sentiments in reports. A last aim was to see what prevalent topics in the reports were and whether these differed by judgment. This paper will not say much about OFSTED, for this I refer to [1].

2. DATA ANALYSIS PROCESS

The procedure that was used for data mining was loosely based on the 'knowledge discovery in data' methodology using CRISP-DM [2]. The Cross Industry standard Process for Data Mining (CRISP-DM) distinguishes several phases that could be applied to the web as well. 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 short paper, and therefore not explicitly mentioned.

2.1 Organizational understanding

OFSTED provides publicly-available inspection reports for every school [3]. Every report has a judgment attached to it which is mentioned on the website and within the report itself. The current judgments are: grade 1 (outstanding), grade 2 (good), grade 3 (requires improvement) and grade 4 (inadequate) [4]. Before

January 2012 grade 3 (requiring improvement) was called ‘satisfactory’ [5]. In addition to the publicly available reports OFSTED also issues interim-reports and letters.

2.2 Data collection and data understanding

A scraper was set up with Scrapy (http://scrapy.org/) and used to scrape the OFSTED website at http://www.ofsted.gov.uk/. The scraper collected the URLs of all historical inspection reports and interim reports since the year of first publication, 2000 (N=9559, 1.39 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 1. The scrape was performed at the beginning of 2014, which explains the lower number of documents for that year. For this paper the publication date of the documents was used over the inspection dates. This was done for two reasons: firstly because we wanted to convey what the ‘outgoing’ message was for the reports, secondly because –as Table 1 indicates– there can be quite some time between these two dates. Just this very scraping process shows that the average number of days between inspection and publication has dropped over the years, with outliers still being quite steep.

Table 1: overview of downloaded OFSTED documents

Year	Reports	Size*)	between inpection & publication			
			Days	Med	Min	Max
2000	212	34.2 MB	171	158	91	1257
2001	278	38.7 MB	101	99	65	347
2002	178	30.5 MB	114	100	79	822
2003	190	33.7 MB	118	99	72	1409
2004	274	35.4 MB	137	94	51	1192
2005	302	51.1 MB	120	73	13	1178
2006	639	58.2 MB	106	26	10	2566
2007	884	122 MB	61	27	7	1975
2008	835	132 MB	42	29	6	1042
2009	896	128 MB	43	30	2	439
2010	1062	150 MB	36	26	10	286
2011	1139	193 MB	39	26	8	800
2012	1000	175 MB	29	22	4	212
2013	1481	239 MB	28	22	9	974
2014	189 **)	7.17 MB	35	31	-1 ***)	120
TOTAL	9559	1.39 GB				

*) Rounded off

**) Up until Feb 15th, 2014

***) This is an error that appeared on the website

Several subsets were created from the collection of documents. The one for this particular paper used all the most recent full inspection reports for every school. Reports were available in PDF format but in preparation of the next phase were converted to txt format. Twenty schools from the 1786 did not yet have an inspection report; most new, like academies, leaving 1766 reports. The 1766 reports were sorted into the different judgments as described in 2.1: inadequate, requiring improvement, satisfactory, good and outstanding. For every judgment the reports were collected chronologically in one folder, with the oldest report first and newest report last.

2.3 Data preparation

In this phase the data were prepared for two different analyses, one involving the sentiment analysis and one the topic modelling.

2.3.1 *Sentiment analysis*

For the sentiment analysis all txt files for a given judgment were chronologically merged into one txt file. The package used in Rstudio Version 0.98.490 [6], R version 3.0.2, was tm.plugin.sentiment [7,8]. The function ‘score’ pre-processes the data by utilizing several functions from the package tm [9,10] by first transforming each text into a ‘corpus’. The three corpora were then subjected to several transformations:

- Making all characters lower case;
- Removing punctuation marks from a text document;
- Removing any numbers from a text document;
- Removing English stop-words;
- Stripping extra whitespace from the documents;
- Stemming the documents;
- Applying a minimum word length of 3;

After this the sentiment models were applied.

2.3.2 *Topic modelling*

For topic modelling all the files in one folder (for one judgment) were transformed in one character vector with each element containing one OFSTED report. Every vector was subsequently tokenized. This was done by removing apostrophes, replacing some characters with space, removing whitespace, making a terms table, tokenizing, removing stop words and not often occurring words. This resulted in the descriptives in Table 2 for Inadequate (Inad), Requires Improvement (Reqi), Satisfactory (Satf), Good and Outstanding (Outs.).

Table 2: descriptives of the topic modelling process for the five corpora

	Inad	**Reqi**	**Satf**	**Good**	**Outs**
Number of documents	149	480	47	854	236
Number of terms	2758	4815	1699	6632	3748
Total number of tokens	301797	908586	98598	1670607	431850

Finally every vector was converted into a list with each element an OFSTED report, so it could be used by the lda package in R [11].

2.4 Modelling

2.4.1 *Sentiment analysis*

At this point five score functions in the tm.plugin.sentiment package were applied to the five judgment documents: polarity, subjectivity, pos_refs_per_ref, neg_refs_per_ref and senti_diffs_per_ref. These sentiment scores are based on the Lydia/Textmap system [12].

- Polarity denotes difference of positive and negative sentiment references divided by the total number of sentiment references.
- Subjectivity denotes the total number of sentiment references divided by the total number of references.
- Pos_refs_per_ref denotes the total number of positive sentiment references divided by the total number of references.
- Neg_refs_per_ref denotes the total number of negative sentiment references divided by the total number of references.
- Senti_diffs_per_ref denotes the difference of positive and negative sentiment references divided by the total number of references.

Table 3 and Figures 2a to 2e present the results for the last of these variables, senti_diffs_per_ref. On the horizontal axis of the graphs IDs for all the documents in the corpora are used, chronologically meaning that ID 1 is the oldest inspection report for that judgment, then ID 2, and so forth.

Table 3: sentiments for the five corpora

Judgment	Senti_diffs_per_ref
Inadequate	0.1757321
Requiring improvement	0.1971455
Satisfactory	0.2062309
Good	0.2046125
Outstanding	0.2116417

Figure 2a: judged inadequate

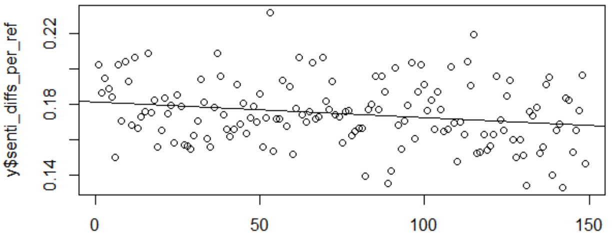


Figure 2b: judged requiring improvement

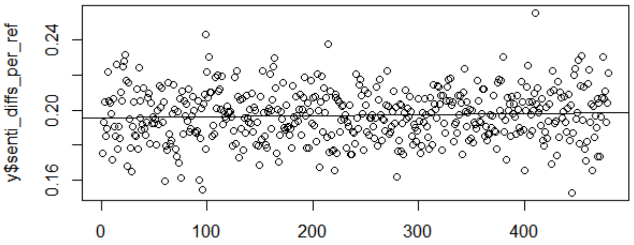


Figure 2c: judged satisfactory

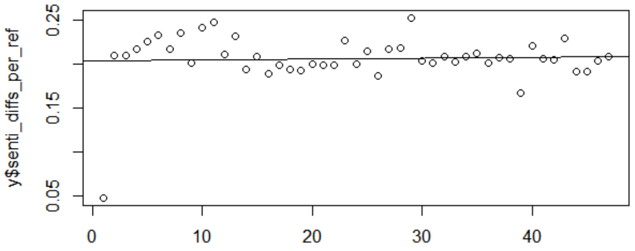


Figure 2d: judged good

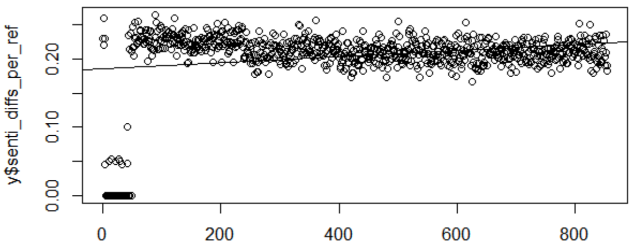
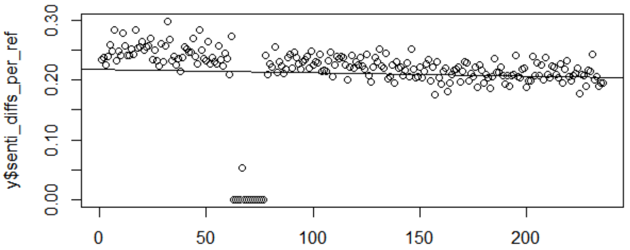


Figure 2e: judged outstanding

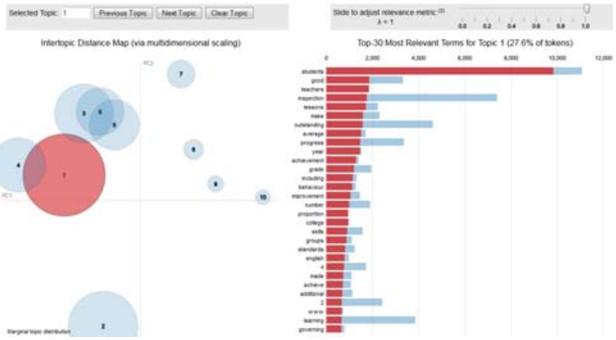


2.4.2 *Topic models*

In machine learning and natural language processing, some statistical models can be used for discovering abstract "topics" in a collection of documents, so-called topic models. One type of topic model can be generated by Latent Dirichlet Allocation (LDA), a generative model which can explain unobserved groups in a set of documents. Each document is a mix of topics and each word in the documents can be attributed to one of the document’s topic [13]. You have a collection of documents with underlying topics. Adopting the approach by Sievert [14], for this dataset LDA was used with the lda package for R and visualizations created with LDAvis [15].

For each of the five judgment corpora a topic model with 20 topics was set up. Priors for the topic-term distributions (eta 0.02) and document-topic distributions (alpha 0.02) were set relatively diffuse. The collapsed Gibbs sampler was set to run for 5,000 iterations. Using the process described in [14] the LDA models were visualized with D3 through the creation of a JSON object. An example of the output for the ‘satisfactory’ category can be seen in figure 3. An overview of the top 5 words for the topic with the most tokens are presented in Table 4.

Figure 3: LDA model visualized with LDAvis [15]



Selected Topic: 1

λ = 1

Table 4: top 5 words for the topic with the largest % of tokens (including the percentage)

Inad	Reqi	Satf	Good	Outs
school	school	schools	good	students
teaching	inspection	satisfactory	inspection	good
inspection	good	lessons	school	teachers
improvement	teaching	pupils	progress	inspection
good	progress	progress	teaching	lessons
42.6%	73.7%	35.3%	35.8%	27.6%

2.5 Evaluation

For the sentiment analysis the following tentative observations can be formulated:

- There seems to be a relationship between OFSTED judgments and sentiments in the inspection reports i.e. reports from schools deemed inadequate have a lower sentiment score than schools judged more favorably. Whether this is a significant difference cannot be concluded from the current analyses.
- The former ‘satisfactory’ judgment which was scrapped [5] seemed to have a sentiment score just around or slightly higher than ‘good’ schools. With the new label ‘requiring improvement’ this is not the case anymore.
- However, over time the reports for inadequate schools seem to have become more negative, while good schools became more positive, closing in on outstanding schools.

For the LDA models it can be observed that the five corpora for the five judgments yield different LDA models with different explanatory power. The topics with the most tokens allocated

3. DISCUSSION

Text mining techniques like sentiment analysis and topic modelling with LDA show promise when it comes to providing a broad indication of the sentiments and topics in sets of documents. This papers shows there is variety in sentiments, as well as word use for different OFSTED judgments. However, notwithstanding this promise, it is important to mention several caveats when applying these techniques, in addition to those described by [16].

Firstly, 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. The multidisciplinary nature of this endeavor on the boundary of both computer science and educational research means that web science is perfectly placed to conduct further research.

Secondly, there also are numerous technical challenges. One issue concerns the processing phase. Even in this experiment the transformation of PDF files was not straightforward. Converting PDF documents to text format depends on whether the file is not password protected. Another challenge concerned special Unicode symbols, missing spaces and other formatting issues like repeating headers. Other choices with regard to choice of stop words, number of topics, scope of the collection, have been made, which might influence the results. It is hypothesized that given the large amount of data and documents the influence was relatively small.

4. REFERENCES

.

[1] Bokhove, C. and Jones, D.K. 2015. Mathematics textbook use in England: mining Ofsted reports for views on textbooks. In Proceedings of the International Conference on Mathematics Textbook Research and Development (Southampton, United Kingdom, July 29-31, 2015). ICMT-2014. University of Southampton, Southampton, 159-166.

[2] Bosnjak, Z., Grljevic, O., and 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

[3] http://www.ofsted.gov.uk/

[4] http://www.ofsted.gov.uk/schools/for-parents-and-carers/understanding-school-inspection-report

[5] http://www.ofsted.gov.uk/news/ofsted-announces-scrapping-of-%E2%80%98satisfactory%E2%80%99-judgement-move-designed-help-improve-education-for-mill

[6] RStudio. 2012. RStudio: Integrated development environment for R [Computer software]. Boston, MA. http://www.rstudio.org/

[7] R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[8] Annau, M. 2013. tm.plugin.sentiment: Text corpus sentiment analysis. R package.

[9] Feinerer, I., and Hornik, K. 2014. tm: Text Mining Package. R package version 0.6. http://CRAN.R-project.org/package=tm

[10] Feinerer, I., Hornik, K., and Meyer, D. 2008. Text mining infrastructure in R. Journal of Statistical Software. 25, 5, 1-54. http://www.jstatsoft.org/v25/i05/.

[11] Chang, J. 2012. Lda: Collapsed Gibbs sampling methods for topic models. R package.

[12] Godbole, N., Srinivasaiah, M. and Skiena, S. 2007. Large-scale sentiment analysis for news and blogs. In Proceedings of the International Conference on Weblogs and Social Media. ICWSM.

[13] Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3, 993-1022.

[14] Sievert, C.P. 2015. A topic model for movie reviews. http://cpsievert.github.io/LDAvis/reviews/reviews.html

[15] Sievert, C., and Shirley, K. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. In ACL Workshop on Interactive Language Learning, Visualization, and Interfaces (Baltimore, Maryland, USA, June 27, 2014). ACL-2014. http://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

[16] Koltcov, S., Koltsova, O., and Nikolenko, S. 2014. Latent dirichlet allocation: stability and applications to studies of user-generated content, In Proceedings of the 2014 ACM conference on Web science (Bloomington, Indiana, USA, June 23-26, 2014). Websci'14. DOI= http://doi.acm.org/10.1145/2615569.2615680.