Institution: London School of Economics and Political Science

Unit of Assessment: 19 – Politics and International Studies

Title of case study: Powerful new tools for text analysis

Period when the underpinning research was undertaken: 2011-2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Kenneth Benoit	Professor of Computational Social Science	2010 to present
Ben Lauderdale	Professor in Research Methodology	2011-2018
Kohei Watanabe	PhD student, later Research Officer	2013-2018
Paul Nulty	Research Officer	2012-2016
Akitaka Matsuo	Research Officer	2016-2019
Adam Obeng	Research Assistant	2016

Period when the claimed impact occurred: 2015-2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact (indicative maximum 100 words)

Research led by Professor Kenneth Benoit at LSE has underpinned the development of new and improved methods of quantitative text analysis and delivered a new, open-access R package (*quanteda*) supporting these. *quanteda* has been used by Facebook and other commercial organisations and data science professionals to improve their products and services, informed international media reporting and analysis, and been widely used in training and development contexts. By 31 December 2020, the *quanteda* package had been downloaded 580,000 times.

2. Underpinning research (indicative maximum 500 words)

Professor Kenneth Benoit's research expertise lies in the development and application of automated, quantitative methods of processing large amounts of "big data", including textual data, and the methodology of text mining. The impacts described here arise from research led by Benoit on the application of novel methods of quantitative text analysis across many disciplines, and the development of new software supporting this **[1] [2]**. The work was conducted at LSE from 2011, primarily as part of a 66-month, EU-funded, ERC Starting Investigator Grant. The research was multi-disciplinary, working with collaborators with backgrounds in statistical analysis and computer simulation, and knowledge of applied domains such as legislative politics.

Quantitative Analysis of Textual Data for Social Sciences (QUANTESS) project

The QUANTESS project (2011-2017, PI Benoit) sought to improve statistical methods for textual data analysis, which hitherto had often relied on untested assumptions and unproven applicability and tended to be based on short "proof-of-concept" demonstrations. This project responded to a lack of substantial academic work explaining the field of textual data analysis for the social sciences. The novelty of QUANTESS lay in its statistical approach to extracting information from texts - treating texts as "data" to be analysed rather than as text to be read and interpreted for "meaning" and categorised or synthesised by humans. It also aimed to develop powerful but accessible free software tools supporting the application of textual data analysis techniques.

QUANTESS led to the creation of a complex and feature-rich software library, enabling users to implement newly developed text analysis methods but also dozens of existing methods, for which there was substantial demand but only limited tools **[3]**. Together, these outputs made up the new *quanteda* package and its companion packages - a library of software functions and data objects allowing user-level programmers to access complex functionality through a simple application-programmer interface **[1]**. *quanteda* is also designed to complement existing packages, to simplify or otherwise enhance aspects of their functionality. To that end, a *quanteda* document feature matrix can easily be parsed to other text-analysis packages for additional analysis or scaling. Benoit's 2018 **[2]** and 2017 **[3]** articles provide overviews of how *quanteda* and related tools can be applied to quantitative text analysis. Because of its development as an open-source platform, during its formative years *quanteda's* analytic techniques were stress-tested many times by knowledgeable online users. Ninety per cent of the 12,000+ lines of code in the software are also covered by unit tests to ensure that their functions behave correctly and robustly.

Since 2017, Benoit has extended the QUANTESS work across several research streams. The first continues to push the methodological frontier of applied text analysis, especially with respect to its applications in the study of political science, rather than in more practical fields such as market research. In a 2019 *American Journal of Political Science* article with Dr Kevin Munger (Pennsylvania State University) and Professor Arthur Spirling (New York University) **[4]**, Benoit presented newly developed quantitative measures of the "readability" of a text (a mechanical computation of difficulty using average sentence length, number of syllables within words, etc.), with an application to measuring the sophistication of political language. This was done using a corpus of State of the Union (SOTU) addresses. The research was able to show that levels of sophistication had indeed lowered over time, consistent with previously voiced concerns over a "dumbing down" of political discourse. One explanation was that the SOTU address had shifted in the mid-20th century to a spoken, rather than written format, broadcast over radio and television and delivered to the nation rather than audiences of politicians in Congress. Findings were reinforced looking at a selection of SOTU addresses to have submitted both a spoken and written version to Congress (seven addresses between 1945 and 1980). The general content was found to be the same but with marked differences in readability, with the spoken address clearly easier to understand than its written counterpart.

Another study, with Dr Alexander Herzog of Clemson University, used the *quanteda* text analysis methods to examine the positions that politicians expressed in their contributions to budget debates between 1987 and 2013, notably in Ireland **[5]**. Analysing voting behaviour was found to be uninformative given that this took place entirely along party lines. By analysing speech, the research showed how politicians, fearful of being punished by their constituents for voting in support of austerity measures, were able to express their opposition in debates, and that such opposition markedly increased in direct response to trends in unemployment in their constituencies or their electoral vulnerability. This analysis was able to reveal an undermining of government cohesion in a way that scrutiny of voting behaviour could not.

In a final stream of the QUANTESS project, the EU-funded "EUENGAGE" project (LSE PI: Benoit) applied *quanteda* software to research on the UK's decision to exit the EU (Brexit), and specifically to tens of millions of social media posts relating to Brexit. This was done using a combination of frequency and keyword analysis, dictionary analysis, and machine learning to classify and compare the language employed by pro-Leave and pro-Remain social media users. Findings revealed significant differences in tone and sentiment in how each group used language **[6]**. Pro-Leave users adopted a more positive tone, using the language of reward, being more oriented towards the future, and more likely to promote assertions of "power" or self-determination. Pro-Remain users, by comparison, adopted a more negative emotional tone, used more "sad" language, were more tentative than assertive, and more oriented towards the past. This research was conducted in collaboration with colleagues at Imperial College London.

3. References to the research (indicative maximum of six references)

[1] Benoit, K. (2019). *quanteda: Quantitative Analysis of Textual Data* (version 1.5.1) [R package]. DOI: 10.5281/zenodo.404692.0. Available at: http://quanteda.io

[2] Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, O., Müller, S., and Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30). DOI: 10.21105/joss.00774.

[3] Welbers, K., van Atteveldt, W., and Benoit, K. (2017). Text analysis in R. *Communications Methods and Measures*, 11(4), pp. 245-265. DOI: 10.1080/19312458.2017.1387238.

[4] Benoit, K., Munger, K., and Spirling, A. (2019). Measuring and Explaining Political Sophistication Through Textual Complexity. *American Journal of Political Science*, 63(2), pp. 491-508. DOI: 10.1111/ajps.12423.

[5] Herzog, A. and Benoit, K. (2015). The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent During Economic Crisis. *Journal of Politics*, 77(4), pp. 1157-1175. DOI: 10.1086/682670.

[6] Matsuo, A. and Benoit, K. (2017). More positive, assertive and forward-looking: how Leave won Twitter. *LSE Brexit* (16 March 2017). Available at: http://eprints.lse.ac.uk/72833/

The research described here received competitive funding and has been published in high-quality, peer-reviewed journals. In 2020, the *quanteda* package won the Society for Political Methodology's prestigious "Best Statistical Software Award".

4. Details of the impact (indicative maximum 750 words)

The underpinning research and its associated software tools have had many diverse impacts. The development of *quanteda* has led directly to the establishment of a new community interest company, been used by commercial organisations, regulators, and data science professionals to improve their products and services, informed media reporting and analysis, and been widely used in training and development contexts, benefitting learning and understanding internationally.

Establishment of new community interest company and subsequent provision of training internationally

The innovation of the *quanteda* R package **[1]** lay in its provision of new tools allowing the implementation of both new text analysis methods developed in QUANTESS, and of dozens of existing methods for which there was substantial demand but only limited tools. To facilitate more efficient dissemination of and engagement with the *quanteda* suite of text-analysis software tools **[1] [2]**, and to ensure their continued development and long-term sustainability, the Quanteda Initiative - a non-profit, community interest company - was established by Benoit and Watanabe in January 2018 **[A]**. Since then, a series of QI-branded training workshops have been held internationally. Up to 31 December 2020, nine training events were delivered for clients in six countries (Germany, Norway, Switzerland, Australia, Japan, and Ireland) **[A]**. Tutorial materials have been prepared in five languages, including Chinese, Japanese, and Hindi. Workshops have typically accommodated groups of around 20 per session. Revenues have subsequently been reinvested into the company to cover the costs of web servers, cloud computing, software licenses, and the services of the three private contractors who have been employed to deliver these expert tutorials.

The *quanteda* suite of text-analysis software tools itself remains completely free to use. And because it is open-source, and therefore completely open to expert scrutiny, the software can be trusted by users. *quanteda* follows a software-as-service model which means that all computation is done on a cloud server; as such, it requires only access to a web browser to use it **[B]**. *quanteda* is designed to be accessible to non-expert user.

Uptake has been significant: by 31 December 2020, it had been downloaded more than 580,000 times **[B]**. This includes by stakeholders in the private sector, by political analysts and commentators at international media organisations, and by data science professionals. User testimonials are available on the *quanteda* website **[C]**. Details of its various impacts and initiatives are included below.

Benefits for commercial users and a communications regulator

In March 2019, Swiss start-up company Grünenfelder Zumbach began using *quanteda*. Grünenfelder Zumbach provide data analytics, policy evaluation, and consulting services for a diverse client base in Switzerland. The speed and stability of the *quanteda* tools have contributed to the more efficient, improved service Grünenfelder Zumbach now provides to its customers, as attested to by a member of its team:

"We are a young company based in Zurich and specialise in policy evaluation. Wherever possible, we try to leverage the potential of data science methods in our projects. quanteda is our first choice when dealing with large amounts of text, because it is fast and stable, it offers many possibilities to quickly find patterns and it provides excellent interfaces for further analysis. Our customers are always impressed [by] how much insight can be gained from large amounts of unstructured text data in short time. Without quanteda our job would be less easy and definitely less fun." **[C]**

Facebook has also used *quanteda*. Its Core Data Science team conducts large-scale, global, quantitative research to gain deeper insights into how people interact with each other and the world around them. The team's findings directly inform improvements to the user experience for Facebook's 2.7 billion monthly active users. A Research Scientist has confirmed the team's use of *quanteda* to conduct its analysis:

"Data scientists at Facebook have used quanteda for a variety of analytic purposes, both as a tool for exploratory analysis and for measurement in relation to production implementation. For example, it has been used to conduct text analysis to measure the linguistic diversity of entity recommendations." **[D]**

Ofcom is the UK's government-approved communications regulator. Ofcom has duties and powers in topics such as media literacy and media plurality, amongst others. Its Economics Group used *quanteda* to conduct analysis during one of its research projects, as confirmed by an Ofcom economist:

"*Economists at Ofcom used quanteda to produce summary statistics, explore possible applications of natural language processing tools, and conduct text analysis – for example it has been used to uncover the topics covered in various documents*." **[E]**

Informing media analysis and reporting

quanteda's text analysis tools have been used to inform media commentary and analysis. In February 2020, i24 news (an Israeli national news and current affairs platform) used *quanteda* in an experimental analysis which compared the texts of two key peace initiatives **[F]**. The first of these was the Peace to Prosperity plan, unveiled by the Trump administration in January 2020 and embraced by Israel; while the second was the Arab Peace Initiative of 2002, put forward by Saudi Arabia and subsequently garnering full Palestinian support. i24's *quanteda*-enabled analysis was able to generate new insights, such as the Trump plan's use of language focused on economic initiatives and its glaring absence of the word "peace", for example.

In July 2020, *The Washington Post* used the *quanteda* Lexicoder Sentiment Dictionary in analysis of whether Russian disinformation campaigns were targeting African Americans **[F]**. This analysis built on a 2018 Senate Select Committee on Intelligence report which concluded that Russian information operatives seeking to interfere with US elections were mainly targeting African Americans. *The Washington Post* used *quanteda* to analyse almost 40,000 purportedly divisive tweets sent from Ghana and Nigeria between June 2017 and March 2020 and believed to be associated with Kremlin-backed sources. Again, the newspaper's *quanteda*-run analysis was able to generate new insights, showing how these malicious accounts tweeted a mixture of sentiments to cultivate followers and manipulate US narratives about race, racial tensions, and police conduct.

Benoit's analysis of the use of Twitter in the run-up to the Brexit referendum **[6]** was used to inform reporting by the UK's *Daily Express*. In January 2017, some six months after the vote, the *Express* quoted comments made by Benoit at an LSE public event and drew upon the EUENGAGE research findings specifically to describe the key impact of more "positive" Brexit campaigners, who were said to have been "*more positive, more certain and more forward-thinking than those supporting Remain, a scientific study of 26 million tweets has revealed*" **[F]**. Similarly, in April 2017 Benoit himself published an article in *The Washington Post* on the question of whether US political discourse is becoming "dumber", in which findings from the State of the Union address analysis **[4]** were discussed extensively.

Impacts on curricula, teaching, and learning

quanteda has provided new tools for teaching in universities around the world, having been added to the curricula of higher education courses in the US and Germany. The University of North Carolina at Charlotte uses *quanteda* in the "Visual Analytics" course of its Data Science and Business Analytics Program **[G]**. In Germany, the University of Frankfurt uses *quanteda* in its Quantitative Text Analysis course **[G]**. Additionally, libraries at higher education institutions have incorporated *quanteda* into their own training and support resources. Notable examples include Princeton University, University of Virginia, WZB Berlin Social Science Centre, and University of Mannheim **[G]**.

Testimonials from *quanteda* users also attest to the software's successful applications to teaching. A user at the University of Bremen describes *quanteda* as: "*an excellent resource for both research and teaching that complements R in a way that is invaluable to me…It is far superior to related packages and so well documented that I use it centrally when teaching text mining*" **[C].** Another user, from Johns Hopkins University (Baltimore, USA), explains: "*I recommend quanteda to students in the Johns Hopkins University Data Science Specialization Capstone project. […]*

quanteda provides a rich set of text analysis features coupled with excellent performance relative to Java-based T packages for text analysis" **[C]**.

quanteda's utility for teaching led to its inclusion in SAGE Publications' Research Methods series, the company's flagship programme of research support resources. Benoit recorded a video explainer for use in "An Introduction to the Quanteda R Package" module, while the software tools have been used in other modules, including "Data Management: Preprocessing with Quanteda" and "Researching Brexit Tweets Using Quanteda Software". *quanteda* also forms part of the syllabus of the SAGE Campus online course, "Fundamentals of Quantitative Text Analysis" **[H]**.

The citation accompanying the 2020 Best Statistical Software Award *quanteda* received from the Society for Political Methodology also expressly mentioned its facility for use in teaching and training. The committee wrote: *"[quanteda's] extraordinary documentation not only makes it accessible for researchers from a variety of backgrounds, it also facilitates the further creation of packages and utilities, and supports its usage in teaching and training… quanteda's innovation, accessible documentation, and functionality are testaments to the collaborative efforts of both junior and senior scholars that can serve as a model for future software development"* **[I]**.

Impacts on data science professionals

In September 2020, *Analytics India Magazine* named *quanteda* as one of its top ten R packages for natural language processing, noting its "*fast, flexible, and comprehensive framework for quantitative text analysis in R*" **[J]**. In October 2019, *Towards Data Science* included the "*all-encompassing*" *quanteda* in its list of the "5 Packages You Should Know for Text Analysis with R", describing it as "*the go-to package for quantitative analysis*" **[J]**. *Towards Data Science* is an online tech journal based in Canada, with 500,000 subscribers.

quanteda has featured on the R-bloggers site, affiliated to the Foundation for Open Access Statistics. Research carried out in 2019 used *quanteda* to analyse 18,000 music reviews posted on the Pitchfork website, with the objective of understanding the linguistic signals of album quality, identifying those adjectives most associated with positive and negative reviews **[J]**.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[A] Quanteda Initiative (quanteda.org), company website featuring details of software, events, and tutorials; established 2018.

[B] *quanteda* (quanteda.io), an R package for managing and analysing text. Features download figures, quick-start guides (in five languages), and associated resources. Data repositories also available at github.com/quanteda.

[C] *quanteda* user testimonials (github.com/quanteda/quanteda/issues/461). GFZB, 7 March 2019; University of Bremen, 25 January 2017; Johns Hopkins University, 18 May 2017.

[D] Supporting statement from Research Scientist, Core Data Science Team, Facebook, 12 December 2020.

[E] Supporting statement from Senior Economist, Ofcom, 29 January 2021.

[F] Examples of use of *quanteda* by media organisations: i24, 10 February 2020; *Washington Post*, 24 July 2020; *Daily Express*, 25 January 2017; and *Washington Post*, 14 April 2017.

[G] Examples of the inclusion of *quanteda* in international teaching courses, including University of North Carolina at Charlotte and University of Frankfurt. *quanteda* also incorproated into training and support resources for Princeton University, University of Virginia, WZB Berlin Social Science Centre, and University of Mannheim

[H] Inclusion of *quanteda* in SAGE Research Methods series modules: "An Introduction to the Quanteda R Package", "Data Management: Preprocessing with Quanteda", and "Researching Brexit Tweets Using Quanteda Software". *quanteda* also forms part of the syllabus of the SAGE Campus online course, "Fundamentals of Quantitative Text Analysis".

[I] 2020 Best Statistical Software Award, Society for Political Methodology.

[J] Examples of data science professionals endorsement of *quanteda*: *Analytics India Magazine*, 7 September 2020; *Towards Data Science*, 6 October 2019; *R-bloggers*, 10 January 2019.