# DATA & computational Journalism

# CONFERENCE

## 1ST/2ND JULY 2019
## UNIVERSITY OF MÁLAGA
## MÁLAGA, SPAIN

CONFERENCE PROCEEDINGS

Google News Initiative · Newtral · Ayuntamiento de Málaga · HACKS/HACKERS SWxSW · HACKS/HACKERS DUBLIN

UNIVERSIDAD DE MÁLAGA · School of Information and Communication Studies · College of Social Sciences and Law · Cardiff University Prifysgol Caerdydd School of Computer Science and Informatics / Ysgol Cyfrifiadureg a Gwybodeg · Cardiff University Prifysgol Caerdydd School of Journalism, Media and Cultural Studies / Ysgol Newyddiaduraeth, y Cyfryngau ac Astudiaethau Diwylliannol · City University of London EST 1894

**Editors:**
Bahareh Heravi, Martin Chorley, Glyn Mottershead



**ISBN:978-1910963388**

**Welcome to the 3rd European Data and Computational Journalism Conference!**

The 3rd European Data and Computational Journalism Conference aims to bring together industry, practitioners and academics in the fields of journalism and news production and information, data, social and computer sciences, facilitating a multidisciplinary discussion on these topics in order to advance research and practice in the broad area of Data and Computational Journalism.

Held in Malaga, Spain, the conference presented a mix of academic talks and keynotes from industry leaders. It was followed by a day of workshops and tutorials.
Submissions of both academic research-focused and industry-focused talks for the conference, on the subjects of journalism, data journalism, and information, data, social and computer sciences were invited for the conference.

**Topics of interest include, but are not limited to:**

- Application of data and computational journalism within newsrooms
- Data driven investigations
- Data storytelling
- Open data for journalism, storytelling, transparency and accountability
- Algorithms, transparency and accountability
- Automated, robot and chatbot journalism
- Newsroom software and tools
- 'Post-fact' journalism and the impact of data
- User experience and interactivity
- Data and Computational Journalism education
- Post-desktop news provision/interaction
- Data mining news sources
- Visualisation and presentation
- News games and gamification of News
- Bias, ethics, transparency and truth in Data Journalism
- Newsroom challenges with respect to data journalism, best practices, success and failure stories

**Collected within these proceedings are the academic abstracts presented at the conference**. We would like to take this opportunity to thank the programme committee for their hard work reviewing submissions and helping us to come up with the fantastic line-up of talks for this year. And an enormous thank you to the organising committee at the University of Malaga for the being such excellent hosts.

Welcome to Malaga, and welcome to DataJConf 2019!

**Bahareh R Heravi, Martin J Chorley & Glyn Mottershead**
DataJConf 2019 co-chairs

For day 2 panels and workshop please visit conference website on datajconf.com.

# Detecting Newsworthy Events in a Journalistic Platform

Tareq Al-Moslmi Marc Gallofrẽ Andreas L OpdahlBjørnar Tessem
Dept. Information Science and Media Studies, Univ of Bergen, N-5020 Bergen, Norway
{Tareq.Al-Moslmi,Marc.Gallofre,Andreas.Opdahl,Bjornar.Tessem}@uib.no

**Abstract:** Social and other open and proprietary data sources are rapidly changing the nature of news and of journalistic work. In the News Angler project, we want to harness such big-data sources for journalistic purposes. We propose a platform, News Hunter, that is able to suggest appropriate news angles on unfolding events to journalists. Precisely assessing the newsworthiness of these events is important to avoid alert fatigue. News angles are seen as patterns that can be matched by news events represented in the knowledge graph. Work on the platform so far suggests that newsworthiness can be estimated as an interplay of at least three factors: reliability – that the event is corroborated by multiple independent and/or trusted sources; match – that the event fits a news angle that is aligned with the intended audience and newsroom profile; and novelty – that the event has not been reported widely from this angle already.

**Keywords:** Journalistic platforms, newsroom systems, knowledge graphs, big data, news angles, news values, newsworthiness.

## Introduction

The News Angler project aims to harness social and other open and proprietary data sources for journalistic purposes. Specifically, we want to leverage the concept of news angles to help journalists effectively identify news events and narrate news stories that may interest their audience. Examples of angles are conflict, local person, and fall from grace. Some angles are more detailed versions of others, such as David-versus-Goliath, a subtype of the conflict angle.

In collaboration with a developer of newsroom systems for the international market, we are developing a platform, News Hunter, that is able to harvest potentially news-relevant text items from the web, analyse them semantically, ingest them into a knowledge graph, aggregate items in the graph into potentially newsworthy events, and suggest suitable news angles on unfolding events to journalists (Berven et al. 2018, Gallofré et al. 2018). News angles are patterns that can match and make interesting events in this knowledge graph. With ever-increasing information, precisely assessing newsworthiness of unfolding events is essential to prevent journalistic alert fatigue.

There is already a broad variety of news-relevant information platforms available (Diakopoulos 2016). They range from general news services such as Google and Yahoo News, through general information platforms such as EMM, OCCRP and WebLyzard, to news-specific ones such as Bloomberg's knowledge graph (Voskarides 2018), Event Registry (Leban 2014) and Reuters Tracer (Liu 2017). Many of them already use knowledge graphs and related semantic technologies, but we are not aware of existing approaches that aim to support news angles and use them to assess newsworthiness.

Our work on the News Hunter platform (Berven et al. 2018, Gallofré et al. 2018) suggests that newsworthiness can be estimated as an interplay of at least three factors: reliability – that the event is corroborated by multiple independent and/or trusted sources; match – that the event fits a news angle that is aligned with the intended audience and newsroom profile; and novelty – that the event has not been reported widely from this angle already.

## Methods

We aim to understand news platforms, news angles, and newsworthiness through design research (Hevner 2007), developing a series of prototypes based on state-of-the-art big data and knowledge graph technologies. Practical relevance is ensured by our industrial partner, who shares their understanding of industrial and journalistic needs and wishes. Theoretical relevance is ensured by focussing on open research issues such as how news platforms can support news angles and how knowledge-graph architectures can scale to big-data settings.

## Findings and Argument

Our work on the News Hunter platform so far suggests that newsworthiness can be estimated as an interplay of at least three factors, which we now discuss in more depth to establish requirements for a news platform that supports angles.

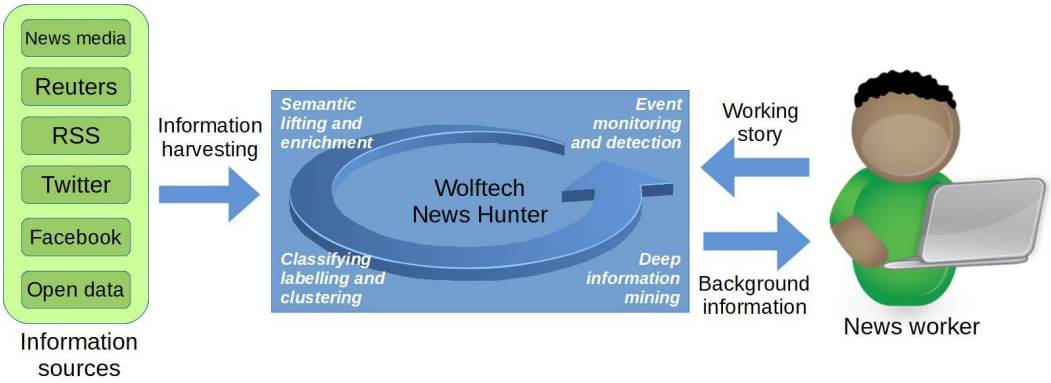


Figure 1 - Overview of the *News Hunter* platform (from Berven et al. 2018).

**Reliability:** Most importantly, in order to be newsworthy, an event must be reliable. If the source is highly trusted, the event may be newsworthy even if it is reported only by a single item. But in most cases, the event must be corroborated by items originating from multiple sources. Reliability of the event is influenced both by the reliability of (or trust in) those sources and of the independence of the items. For example, two tweets may be based on the same underlying source or one may simply be a retweet of the other. Finally, event reliability is also influenced by how reliable the lifting of textual items into semantic item graphs and the aggregation of those item graphs into event graphs were. To support corroboration, the news platform must therefore be able to trace events back to their originating items and those items' sources. Trust in sources must be estimated and maintained, as well as trust in the lifting and aggregation steps on the way from textual items to event graphs. To the extent possible, the external sources of items should also be identified – at least items that are based directly on one another or on a common precursor need to be identified.

**Match:** In order to be newsworthy, the event must match a news angle, which is a pattern for moulding an event, if possible, into a fabula, which is a sub-graph of facts about the event that can be narrated to become a story. It is important that angles do not only fit the event, but also the news organisation's intended audience and profile. To support angle matching, the news platform must therefore maintain a library of angles, whether created manually or learned automatically. It must be aware of which events and angles that fit the audience and newsroom profile, and it must be able to match angles with events to mould fabulas. Also, to match a news angle, the event graph must be sufficiently detailed. This is another reason for aggregating news items into event graphs, which will presumedly not only be more reliable, but also more detailed and complete than the individual item graphs. The news platform should invite the journalist to collect further facts when needed to complete a promising angle.

**Novelty:** Finally, in order to be newsworthy, an angled event should be original. Other news media that target the same audience should not already have covered the event from the same angle. To support novelty, the news platform must therefore harvest news items published by competing media organisations that target a similar audience in real time. It must be able to trace from a news item to the event it describes, and it must be able to detect the angle from which an event is narrated.

## Conclusions

We have presented the objectives of the News Angler project and how we plan to identify newsworthy events by assessing their reliability, match, and originality. In future work we will validate our approach by continuing to extend the *News Hunter* platform to support newsworthiness and news angles.

## References


Berven, Arne, et al. (2018) “News Hunter: Building and mining knowledge graphs for newsroom systems.” Proc. NOKOBIT 26, Svalbard.

Diakopoulos, Nicholas (2016) "Computational journalism and the emergence of news platforms." Routledge Comp. Dig. Journalism Studies.

Gallofré Ocana, Marc, et al. (2018) "Towards a Big Data Platform for News Angles⋆." Proc. NOBIDS’18, Trondheim.

Hevner, Alan R. (2007) "A three cycle view of design science research." SJIS 19.2:4.

Leban, Gregor, et al. (2014) "Event registry: learning about world events from news." Proc. 23rd Int. Conf. WWW.

Liu, Xiaomo, et al. (2017) "Reuters tracer: Toward automated news production using large scale social media data." IEEE Int. Conf. Big Data.

Voskarides, Nikos, et al. (2018) "Weakly-supervised Contextualization of Knowledge Graph Facts." Proc. 41st ACM Int. SIGIR Conf.

# Fake News Detection Based on Named Entity Recognition and Machine Learning

Francisco L. Valverde
Dept. of Computer Science University of Malaga
valverde@uma.es

Rafaela Benítez Rochel
Dept. of Computer Science University of Malaga
rbenitezr@uma.es

María Guerrero Aguilar
Journalist of Press Office University of Malaga
mariaguerrero@uma.es

**Abstract:** False news has become a problem of the first magnitude for the governments of nations and the media. Due to the large volume of information to analyze to solve this problem it seems that the solution should be an automatic method that manages to detect false news. However, today we still do not have the technology to have an automatic and efficient solution. For this reason, the only solutions that are working are based on manual operation. Our proposal consists of a decision support system for these organizations to facilitate their work. Using a machine learning system based on Named Entity Recognition and identities of the authors it is possible to make a prior classification of authenticity. Thanks to this scheme, the amount of information that is necessary to analyze manually is greatly reduced.

**Keywords:** Fake News Detection, Named Entity Recognition, Machine Learning. Super Vector Machines, Identity authenticity.

## Introduction

In this paper we present a proposal for a semiautomatic scheme for the detection of false news. Over recent years, the extensive growth in the number and types of fake news has led to the necessity for building and effective detection system for fakes news identification with the capability of handling the volume, the variety and the velocity associated with them. Augey and Alcaráz (2019) in a recent investigation conclude that most of the false news is motivated by financial objectives. We are, therefore, faced with a new challenge that, as Mc Nair (2017) points out, does not respond to an isolated cultural problem, but to the result of the social trends of the 21st century. Globalization, the rise of relativism, the crisis of objectivity, the consumption of digital news or the decline of trust in journalism are some of the factors that this author identifies as explosives of the rise of false news. In relation to the online fake news audience, the latest works indicate that it is a subset of the total, disloyal and high availability news audience on the Internet (L. Nelson & Taneja, 2018).

The World Economic Forum (WEF) has been warning for years about the global danger of the massive existing digital disinformation, as a technological and geopolitical risk. Likewise, the European Barometer 464 on 'False news and disinformation on line', made in 2018, detected a high degree of belief in Spain of being exposed to fake news. Its direct consequences in politics are also the subject of many other studies. In this line, an article published in the research center 'Pew Research Center' (2016) claimed that most Americans (64 percent) suspected that the false news generated confusion and had a potential impact on both political life as in the individual citizens. J, Vargo and A. Amazeen (2017) warn of the relationship of these news with the digital partisan media, which they identify as highly sensitive, however they downplay their impact on the new media, although these also respond to the false news agendas.

In this work we propose a decision support system that facilitates the work of agencies that specialize in the detection of false news. Using a model of machine learning based on Named Entity Recognition (NER) and identity authenticity, it is possible to make a massive preliminary classification with an accuracy of more than 80 percent

## Methods

Currently, there are facts checking organizations such as Snopes, Politifact, TruthOrFiction, Factcheck, OpenSources, FakeNewsWatch, fakespot, reviewmeta, Opensecrets.org ,etc which operate on the basis of the traditional journalistic model. In these organizations the reporters have to evaluate facts in order to obtain the veracity of a statement. This approach is not automated and is often time-consuming and difficult to compete with the quantity of fake news published daily. This problem has led researchers and technical developers to look at several automated ways of assessing the truth value of potentially deceptive text based on the properties of the content and the patterns of computer-mediated communication.

Machine learning: Supervised machine learning algorithms like Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, K-nearest Neighbour are extensively used in previous literatures for online hoaxes, frauds, and deceptive in- formation classification (Afroz et al., 2012); deep learning based methods are good solutions for online fake news representation and detection, and have been introduced in Ruchansky et al. (2017). Unsupervised learning model for fake news detection, they are: cluster analysis , outlier analysis, semantic similarity analysis (Li, McLean, Bandar, O'shea, & Crockett, 2006), and unsupervised news embedding techniques include Word2vec, FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), Sent2vec (Pagliardini, Gupta, & Jaggi, 2017), and Doc2vec (Le & Mikolov, 2014). Our method is based on the Named Entity Recognition on news and identity of authors.
Using only this information in a Super Vector Machine it is possible to classify a story as probably authentic or probably false. Our focus is on the simplicity of the analysis that makes it more appropriate when analyzing large amounts of news. Only those news that have been classified as false are analyzed manually by the organizations to ensure and confirm this classification.

## Findings and Argument

On a simulated data set of 100 false news and 100 truthful news, the SVM system was able to correctly detect and classify 81.5% of the news. As you can see in figure 1 the classification is quite accurate.

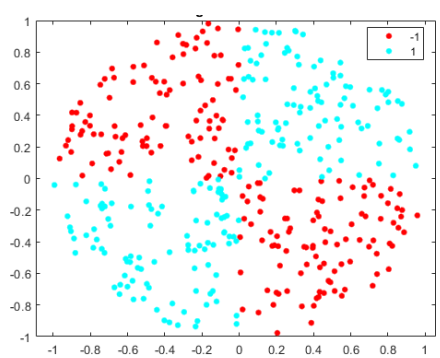


*Figure 1 – News classification by SVM based on NER. 1 indicates true and -1 false*

## Conclusions

A new hybrid method to detect false news has been presented in this article. The main contribution is the simplicity that makes it suitable for analyzing large amounts of data. It is a decision support system for companies specialized in the detection of false news since it facilitates their work and greatly reduces the speed and costs.

## References


Alcaraz, M. & Augey, D. (2019) Will Fake News Kill Information? In book: Digital Information Ecosystems. Journal of Communication, pp 139-159. DOI: 10.1002/9781119579717.ch7

Afroz, S., Brennan, M., & Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. Security and privacy (sp), 2012 IEEE symposium on.IEEE461–475.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.

Barthel, M., Mitchell, A. & Holcomb, J. (2016). Many americans believe fake news is sowing confusion. Pew Research Center. Retrieved from: https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

European Commission (2018). Eurobarometer 464 on "Fake News and Disinformation on line". Retrieved from: http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/instruments/flash/surveyky/2183

Gualda, E. (2019) Teorías de la conspiración, confianza y credibilidad en la información. Communication & Society VOL. 32(1). Retrieved from: https://www.unav.es/fcom/communication-society/es/articulo.php?art_id=728

Hernández, A. (2017). Resiliencia de la organización de la información en la era de las posverdad. Alcance, 6(14), 47-59.

High-Level Expert Group on Fake news and Disinformation (2018). A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation. March. European Commission, Directorate-General for Communication Networks, Content and Technology, Luxemburg. Retrieved from: https://publications.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. International conference on machine learning1188–1196

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8), 1138–1150

McNair B. (2017) Fake News. Falsehood, Fabrication and Fantasy in Journalism. Routledge Focus. 1st edition.

Pagliardini, M., Gupta, P., & Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. arXiv:1703.02507

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on conference on information and knowledge management. ACM797–806.

Vargo, J., Guo,L. & Amazeen, A. (2017) The agenda-setting power of fake news: A big data analysis of the online landscape from 2014 to 2016. New Media & Society, vol. 20, 5: pp. 2028-2049.

Weir, W. (2009). History´s Greatest Lies. Beverly: Fair Winds Press.

# RODA: a tool for semi-automatic data-driven visual stories

Xaquín Veira-González
Universitat de Girona, Graphics and Imaging Laboratory
xocasgv@gmail.com

Anton Bardera
Universitat de Girona, Graphics and Imaging Laboratory
anton.bardera@imae.udg.edu

Apple Chan-Fardel
apple.cjcfardel@yahoo.com

María Luisa Otero López Universidade de Santiago de Compostela
marotlo@usc.s

**Abstract:** RODA, our robot data assistant, is an interactive tool for data inquiry and automation of data-driven visual narratives. So far, most of the research on data-driven automation in journalism has drilled down on either text narratives or isolated data visualizations. Our research for RODA is the first time that the aim is to produce narratives incorporating semantically interwoven text and visualizations. By using the current advances in natural-language generation and research on structures of this type of storytelling devices, we expand the theoretical framework of data-driven visual storytelling.

**Keywords:** data-driven storytelling, information visualization, natural-language generation, narrative visualization, narrative automation, data-driven visual stories, robot journalism, data journalism, visual storytelling

## Introduction

Leading newsrooms are devoting more resources to make visual-driven formats integral to their vocabulary. The New York Times' 2020 report "Journalism That Stands Apart" charted the growth of "stories with deliberately placed visual elements" from close to nothing in 2014 to 12.1% by September 2016. "Deliberately placed" is the crucial nuance that hints to a holistic way of editing, as Cairo (2017) describes it, that is currently being used in many of leading newsrooms: text, visualizations, pictures, and videos are just different media types to tell a portion of a story, which flows in and out of them seamlessly. Text and graphics blend in the narrative; they use short sentences anchored in summary statistics that refer to what the graphics show.

Many tools automate the production of charts from data, and the current advances in NLG are starting to provide tools to automate the writing of stories from data. We present a robot data assistant (RODA) with the ambition to automate the production of complete narratives that coherently mix text and visualizations based on the input data. The user would enter a dataset into the system, and following a conversation with the application the robot would try to understand the data, question the user about necessary checks, summarize possible patterns or trends, recommend visualization types, and use natural-language generation (NLG) to assemble a narrative with text and visuals that would fit the priorities of the communicator.

## Methods

The focus of information visualization (InfoVis) had been, until recently, on interactive visual representations of data as isolated interfaces for the data. Segel and Heer's (2010) shifted the focus and introduced the concept of narrative visualizations, as a combination of visuals, multimedia and textual elements integrated within data-driven storytelling systems. In Riche et al. (2018), practitioners and researchers explore storytelling techniques, the lifecycle of the story, and narrative patterns, defining future lines of research and exploration for data-driven visual storytelling. Following the patterns in Veira-González and Perez-Montoro's (2018) we explore the underlying structures of these data-driven visual stories and their atomic components.

## Findings and Argument

RODA is designed as a blend of chatbot and on-screen interaction with the user from which it calculates the summary statistics, recommends charts, understands what the data means, gathers the user's priorities for the story, and outputs a story structure composed of narrative blocks with semantically interconnected text and visuals.

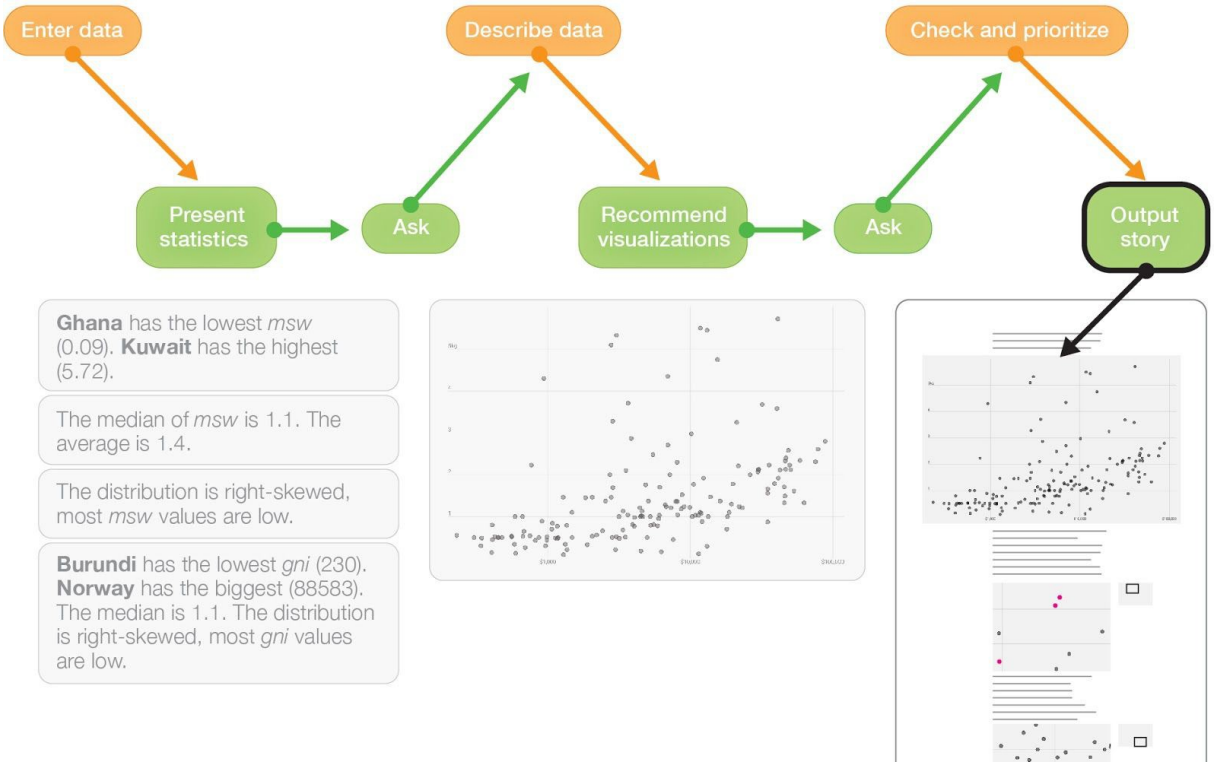


*Figure 1 - Streamlined flow chart of how RODA's inputs and feedback to the user work*

Each story atom is composed of a data description text, a visualization, and an explanation and transitional text all tied to the current view of the data. They resemble Kosara's (2017) Claim, Fact, and Conclusion (CFO) pattern based on Cohn's (2013) narrative structure for comics.

RODA's recommendations rest on research that two of the authors did while designing The Guardian's in-house charting tool. While the implementation starts with a dataset and then suggests a visual display, our approach walked backward from a dozen of visualizations: the system parses the data types of the input, the ranges of the numeric properties, and filters the available visualization methods based on a set of constraints.

Around each instance of the visualization, the text refers to the critical features visualized, and if it is known, provides the context and the reasons for what it is salient. Text blocks in these type of stories, especially the explanatory copy, serve a similar function to the annotation layer within data visualizations.

In order to automatically detect facts from the data for content planning of the transitional text, several statistics have to be computed. In the first iteration of the prototype, we focus on basic statistical measures, such as mean, median, percentiles and quantiles, and standard deviation. The tool uses rule-based generation to write the text, a more sophisticated method than fill-in-the-blanks-with-data templates.

The user's answers determine the content for explanatory texts in the final part of the application flow. Once the tool has computed a set of facts and has understood how to refer to the data, it can ask to add context or reasons behind some of those calculated facts, such as "do the outliers have anything in common," "how about items near the average or the median." The application would then summarize those answers. In order to structure those narrative blocks and compose the story, the approach that best fits our purposes is Kosara's (2017) Claim, Fact, and Conclusion (CFO) pattern, based on Cohn's (2013) narrative structure for comics. We also use Veira-González and Perez-Montoro's (2018) story patterns to determine the order and scope (overview or details) of the narrative blocks.

## Conclusions

What RODA can do in its current iteration is inherent to its purpose. It isn't just a tool for automating data-driven stories, but almost more importantly a tool for training journalists: an aid to contribute to data and visual literacy in newsrooms and other communication environments.

Some of the limitations also come from the relatively small body of research on the data-driven visual stories —still in its infancy compared to other subfields of InfoVis— and the fact that this paper is a first-ever approach to automating these type of narratives.

Future research will gain as well from surveying the effectiveness of these machine-written visual stories compared both to human-written text pieces and to individual charts by themselves.

## References

Cairo, A., 2017. Nerd journalism: How data and digital technology transformed news graphics (Doctoral dissertation, Universitat Oberta de Catalunya).
Cohn, N., 2013. Visual narrative structure. Cognitive science, 37(3), pp.413-452.
Kosara, R., 2017, June. An argument structure for data stories. In Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers (pp. 31-35). Eurographics Association.
Pérez-Montoro, M. and Veira-González, X., 2018. Information Visualization in Digital News Media. In Interaction in Digital News Media (pp. 33-53). Palgrave Macmillan, Cham.
Riche, N.H., Hurter, C., Diakopoulos, N. and Carpendale, S. eds., 2018. Data-driven storytelling. CRC Press.

# Becoming a data journalist: the role of identity in journalism education

Liz Hannaford Manchester Metropolitan University
L.Hannaford@mmu.ac.uk

**Abstract**: As data journalism becomes mainstream, journalism educators need to find ways to bring it in to their teaching. However, the literature shows that this can be problematic for staff and students in an already tightly-packed curriculum. The objective of this study is to explore ways in which learning to do data journalism can be reconceptualised as a social process of becoming a data journalist whereby students are invited to take on the beliefs and values of the new professional identities made desirable in a datified society.
I address this problem by using a discourse analysis approach to explore how vocal advocates in this field use different discursive strategies to justify their practices. Preliminary findings suggest speakers redraw the boundaries of journalism as they negotiate and lay claim to competing identities with implications for journalism educators.

**Keywords:** Data journalism education, identity, Community of Practice, discourse analysis

## Introduction

This paper sets out the background, rationale, methods and some initial findings of my doctoral research, currently ongoing, which is a study of the discourses of data journalism, the construction of data journalist identities and the implications for journalism education. Encountering spreadsheets, statistics and code can be a jarring experience for undergraduates who did not expect their journalism course to be 'technical.' Educators need to address what professional and social identities we are inviting students to invest in when they study journalism and how these identities have been recalibrated by journalism's 'quantitative turn.' To achieve these aims, my research is analysing the taken-for- granted norms that have become embedded in the discourse of data journalism, and asking whether and how this discourse can be exclusionary.

There is now a small but growing number of specialist data journalism courses at Masters level in the UK (Bradshaw, 2018), but attempts to introduce data journalism skills into traditional journalism programmes have encountered obstacles. These include journalism students' aversion to maths, a fear that students are put off by the subject and the lack of qualified staff to teach data journalism (Hewett, 2015).

The literature on data journalism education has predominantly focused on the extent to which it is taught around the world and the challenges it presents (Splendore et al., 2016 in six European countries; Berret and Phillips, 2016 in the United States; Yang and Du, 2016 in Hong Kong; Davies and Cullen, 2016 in Australia; Heravi, 2019 globally). Elsewhere, research has explored the importance in this field of peer-to-peer learning through informal networks such as the NICAR listserv (Howard, 2014; Fink and Anderson, 2015; Hermida and Young, 2017), HacksHackers meet-ups around the world (Lewis and Usher, 2014) and dedicated social media groups (Appelgren, 2016), all of which suggest the importance of social identity and community participation in this field.

The implications of these learning practices for formal journalism education have yet to be fully explored in the literature. Given the need for journalism education to embrace basic data skills as a requirement of the profession (Stalph and Borges-Rey, 2018), it is argued here that socio- cultural perspectives could provide valuable insights into learning as an ongoing social process. Research from other professions provides support for this approach (Monrouxe, (2010) in medical students and Beauchamp and Thomas, (2009) in student teachers, for example).
This paper presents initial findings from analysis of the discourses of data journalism as deployed by its influential early pioneers during the social interaction of interviews and panel discussions. The analysis suggests that the dominant discourses of data journalism often rely on negative representations of traditional journalism as 'broken' whilst representing technologists as heroic saviours.
These discursive strategies talk data journalism 'out' of the journalism curriculum and can antagonise students' social identity. I argue that it would be more beneficial to find ways of talking it 'in' to the curriculum to help students manage the transition to the new professional identities required in a datified society.

## Methods

The methodology is driven by the following research questions. How do vocal advocates of data journalism talk about this field? What identities (subject positionings) and practices do these ways of talking make possible? What are the implications for journalism educators?

The research is currently ongoing. The preliminary analysis presented in this paper is based on using a discourse analysis approach (Jager and Maier, 2016), which builds on recent interest in investigating data journalism as a socio-discursive construct produced through social interaction (Powers, 2012; De Maeyer et al., 2015; Borges-Rey, 2017) as opposed to an inevitable reality existing 'out there'.

The data to which this analysis has been applied consists of transcripts of interviews and panel discussions involving prominent practitioners and pioneers in North American and European data journalism from 2008 to 2018. To produce a manageable data set for a discourse analysis approach, a purposive selection of interactional texts was made. The advantage of analysing social interaction texts as opposed to written texts is that they are a rich source of narratives about how the speakers see themselves and what they do.

## Findings and Argument

The research is ongoing, and as such the findings presented are preliminary. A number of different discursive strategies were used by vocal advocates of data journalism to legitimise their practice. There is evidence of the use of negative representations of traditional journalism as 'broken', the idealised pursuit of journalistic 'truth', the unquestioned privileging of numerical, structured data over other forms of knowing, a radical vision of the future of journalism but also positive representations of technology as emotionally fulfilling and a discourse of optimism about journalism's future.

These discourses involve the speakers negotiating and laying claim to different and competing identities as they redraw the boundaries of journalism and explore new ways of becoming a journalist. These negotiations take place against the background of an existential crisis in journalism. Neo-liberal discourses thus blend with these new identities as journalism becomes a digitised commodity in a global market. Practices required by these ways of talking about data journalism include the fetishisation of openness, collaboration, disruption and innovation. Continual learning – often from peers – is highly valued and aligned with social participation in communities of likeminded practitioners that transcend organisational boundaries.

## Conclusions

The preliminary conclusions of the research suggest that journalism's quantitative turn requires more from educators than just squeezing new skills into an already tightly-packed curriculum. I argue that knowledge and identity are intertwined (Lave and Wenger, 1991) and so educators need to consider who students need to be as much as what they need to know. The experience of becoming a journalist repeatedly raises issues of identity, values and beliefs that need to be addressed in the classroom. Students have to be able to make sense of

themselves in this new journalism-technology environment and, if they choose, inhabit the professional identities that result.

## References


Appelgren, E. (2016) 'Data Journalists Using Facebook.' Nordicom Review, 37(1) pp. 156-169.

Beauchamp, C. and Thomas, L. (2009) 'Understanding teacher identity: an overview of issues in the literature and implications for teacher education.' Cambridge Journal of Education, 39(2) pp. 175-189.

Berret, C. and Phillips, C. (2016) Teaching Data and Computational Journalism. New York, NY: Columbia Journalism School.

Borges-Rey, E. (2017) 'Towards an epistemology of data journalism in the devolved nations of the United Kingdom: Changes and continuities in materiality, performativity and reflexivity.' Journalism (published online ahead of print 1st February) Available at DOI: 10.1177/1464884917693864

Bradshaw, P. (2018) 'Data Journalism Teaching, Fast and Slow.' Asia Pacific Media Educator, 28(1) pp. 55-66.

Davies, K. and Cullen, T. (2016) 'Data Journalism Classes in Australian Universities: Educators Describe Progress to Date.' Asia Pacific Media Educator, 26(2) pp. 132-147.

De Maeyer, J., Libert, M., Domingo, D., Heinderyckx, F. and Le Cam, F. (2015) 'Waiting for Data Journalism: A qualitative assessment of the anecdotal take-up of data journalism in French-speaking Belgium.' Digital Journalism, 3(3) pp. 432-446.

Fink, K. and Anderson, C. (2015) 'Data Journalism in the United States: Beyond the “usual suspects”.' Journalism Studies, 16(4) pp. 467-481. Jäger, S. and Maier, F., 2016. Analysing discourses and dispositives: a Foucauldian approach to theory and methodology. In Wodak, R. and

Meyer, M. (eds) 3rd ed. Methods of critical discourse studies, Los Angeles: Sage, pp.109-136.

Heravi, B.R., 2019. 3WS of Data Journalism Education: What, where and who?. Journalism Practice, 13(3), pp.349-366.

Hermida, A. and Young, M. L. (2017) 'Finding the Data Unicorn: A hierarchy of hybridity in data and computational journalism.' Digital Journalism, 5(2) pp. 159-176.

Hewett, J. (2015) 'Learning to teach data journalism: Innovation, influence and constraints.' Journalism, 17(1) pp. 119-137.

Howard, A. B. (2014) The Art and Science of Data-driven Journalism. New York: Tow Center for Digital Journalism. [Online] [Accessed on 21st March 2019] Available at https://doi.org/10.7916/D8Q531V1

Lave, J. and Wenger, E. (1991) Situated learning: Legitimate peripheral participation. Cambridge University Press.

Lewis, S. C. and Usher, N. (2014) 'Code, Collaboration, And The Future Of Journalism: A case study of the Hacks/Hackers global network.' Digital Journalism, 2(3) pp. 383-393.

Monrouxe, L. V. (2010) 'Identity, identification and medical education: why should we care?' Medical education, 44(1) pp. 40-49.

Powers, M. (2012) '“In Forms That Are Familiar and Yet-to-Be Invented” American Journalism and the Discourse of Technologically Specific Work.' Journal of Communication Inquiry, 36(1) pp. 24-43.

Splendore, S., Di Salvo, P., Eberwein, T., Groenhart, H., Kus, M. and Porlezza, C. (2016) 'Educational strategies in data journalism: A comparative study of six European countries.' Journalism, 17(1) pp. 138-152.

Stalph, F. and Borges-Rey, E. (2018) 'Data Journalism Sustainability: An outlook on the future of data-driven reporting.' Digital Journalism, 6(8) pp. 1078-1089.

Yang, F. and Du, Y. R. (2016) 'Storytelling in the Age of Big Data: Hong Kong Students’ Readiness and Attitude towards Data Journalism.' Asia Pacific Media Educator, 26(2) pp. 148-162.

# Predictive sentiment analysis of messages for Journalistic Purposes: Real-time classification of tweets based on Machine Learning

Prof. Dr. Félix Ortega
Universidad de Salamanca
E-mail fortega@usal.es

Dr.Carlos Arcila
Universidad de Salamanca
E-mail carcila@usal.es

Prof. Antonio García
Universidad Rey Juan Carlos
E-mail antonio.garcia@urjc.es

**Abstract:** Algorithms, big data, machine learning and artificial intelligence systems are key concepts and methods for the reshaping of our sociocultural, economic and political relations in our everyday life. Digital culture and communication are inevitably changing as media infrastructures, media practices and social environments become increasingly more data conscious-driven. The efforts in bringing together automated sentiment analysis based on machine learning and streaming technologies that produce important amount of data, are relatively new to journalistic oriented media enterprises. This paper describes and assess the creation of machine learning models to predict sentiments in real- time tweets associated to the under revolution Journalistic Value Chain (JVC) and depicts how this process can be scaled using commercial distributed computing when personal computers do not support computations and storage in order to provide the Data Journalistic Unit (DJU) with tools for a better job and information performance.

**Keywords:** Predictive sentiment analysis; Political Opinion; Twitter; Machine Learning; Big Data; Political tweets; Big data digital journalism.

## Introduction

Algorithms, big data, machine learning and artificial intelligence systems are key concepts and methods for the reshaping of our sociocultural, economic and political relations in our everyday life. Digital culture and communication are inevitably changing as media infrastructures, media practices and social environments become increasingly more data conscious-driven. The consumer´s use of the commonplace media technologies in a world of information and news is mediated by data, interpelatting and adapting to consumer preferences in an everyday more automated way. We live in a world which is increasingly influenced by algorithms and artificial intelligence methods and processes. This trend is now spreading rapidly to the analysis and comprehension of the flow of information, opinions and news in the digital media. Algorithms, machine learning and artificial intelligence is being implemented in the filtering of a large percentage of the content published on social media platforms and their Apps, picking out what is potentially newsworthy (Thurman et al., 2016, 2017) for the consumer given its preferences and transforming news` management and agenda setting in media enterprises into in a more Data Broker Management Journalism (DBMJ) where new journalistic competences converge with online and almost real time analysis of the “trends”, “visits”, “ratios,… and broker marketing oriented visualizations.

In this context, there is a growing interest in surveying opinions using large-scale data produced by social media (Cobb, 2015; O'Connor, 2010; Bollen, Mao & Pepe, 2011) in media enterprises and in particular the traditional journalistically oriented business. The vast majority of these research is based upon either on manual classification or automated content analysis using dictionaries that score words (e.g. giving an a priori negative or positive value to each word) (Leetaru, 2012; Feldman, 2013) and other approaches such as supervised machine learning (Vinodhini & Chandrasekaran, 2012) are scarce in communication research (van Zoonen & Toni 2016). Moreover, the efforts in bringing together automated sentiment analysis based on machine learning and streaming technologies that produce important amount of data, are relatively new to journalistic oriented media enterprises. This paper describes and assess the creation of machine learning models to predict sentiments in real-time tweets associated to the under revolution Journalistic Value Chain (JVC) and depicts how this process can be scaled using commercial

distributed computing when personal computers do not support computations and storage in order to provide the Data Journalistic Unit (DJU) with tools for a better job and information performance.

Local computing solutions for journalistic purposes may provide serious limitations in the range of high amount of data analysis which necessarily requires scalable storage and distributed computing. Running streaming data analysis in distributed platforms has been challenging in the complex and changing big data landscape (Turck & Hao, 2016). The incorporation of tools such as Apache Kafka has allowed the current most extended open software for distributed computing Apache Spark to fulfill this gap with Spark Streaming (Spark Kafka Integration, 2016), which can read code in Scala or also in Python (with the module PySpark). We analyze in our research how journalists can extend sentiment analysis with Apache Spark Streaming in local machines using trained models with Spark Machine Learning. We also explain how this procedure is scalable using commercial tools (instead of academic grids) such as the most popular Infrastructure as a Service (IaaS) Amazon Web Services (AWS), that offers Amazon S3 for massive storage and Amazon Elastic Computing Cloud (EC2) to create a flexible set of connected instances in the cloud in order to compute the analysis.

## Methods

The computational methods and services explained in our research may contribute to help journalists in media enterprises study, interpret and analyze big amounts of tweets in any language running sentiment analysis in real-time. These methods and techniques do require some programming skills, however exiting models allows short-time learning curves for the final user providing easy adaptations. We provide journalists with all the code for Spark (written in Python and using PySpark) in a iNoteBook (ipynb). No mathematical background is needed to run the machine learning models, but a theoretical understanding of the algorithms will increase the quality of the interpretation for the trained journalist. In the case of the mentioned commercial services (AWS, Azure, IBM, etc.) media enterprises must consider the financial costs associated for this analysis. In addition, working with interdisciplinary teams (computer scientists, statisticians, computational linguistics, etc.) can improve the results and save resources for the design of the Data journalistic enterprise. The described procedure to monitor tweets in streaming might help testing traditional and emerging theoretical approaches in communication research that require longitudinal data and might also contribute to experimental studies which need real-time inputs to create or adapt to stimuli, understanding the media journalistic value chain in a reciprocal way is key to the consolidation of the professional profiles of the Data Analysis Era.

## Findings and Argument

In a sector associated to innovation and technology, media enterprises are adapting and transforming their workflow structures into a more automated journalistic value chain, where artificial intelligence software is substituting traditional journalistic “roles” into a scenario where little or no human intervention is required aside from the software-hardware implementation and programming (Carlson, 2015) in some specific niche news production and redistribution like social networks. If I may use a metaphor to illustrate, robotics ongoing implementation is revolutionizing the automobile industry at a continuous pace, likewise the development of big data broadly, artificial intelligence processes and machine-learning and written news is opening a new scenario where technology providers implement algorithms and artificial intelligence processes to deliver automated news in multiple languages with ethical challenges arising (Dörr et al, 2016, 2017). This is the true revolution for journalistic-media enterprises. Algorithms are being used in new ways to distribute and package news content, both enabling consumers to request more of what they like and less of what they don’t and also making decisions on consumers’ behalf based on their data preference curve and profile (Groot Kormelink and Costera Meijer, 2014). We provide journalists as indicated above with all the code for Spark (written in Python and using PySpark) in a iNoteBook (ipynb). No mathematical background is needed to run the machine learning models, but a theoretical understanding of the algorithms will increase the quality of the interpretation for the trained journalist.

## Conclusions

The social role of journalism will prevail as a longstanding facilitator and interpreter of what is going on, but the labour process of the journalistic roles and authorities will merge into an Artificial Intelligence role and a human based value added provider. The quality of the news, their interpretation, the final accountability will be placed progressively in the hands of artificial neural nets and human neurons simultaneously. As implemented in the automobile industry, the human neurons will supervise and programme the software-hardware-robotic processes work flow and contribute where complex reasoning and reprogramming, or complex writing is at need and socio-economic and politically viable. There should not be a catastrophic concern about the quality of the news, their transparency and accountability with the data revolution in place. It will remain almost as it is today and improving with more data available and analysis. It will imply less human intervention where "machine learning processes" will prevail given their comparative advantages to that of the human workforce. The journalistic worker will evolve into a Data Broker Management Journalist (DBMJ) riding his "data driven surf board" in an all digital revolutionised Journalistic Value Chain (JVC), following data and news labelled via blockchains for better traceability.

This renewed journalist will adjust processes and take decisions by supervising the job done by the Artificial Intelligence Big Data News solutions, filtering and contrasting "fake or irrelevant" news and in some more value-added and quality-orientated media business will to some extent complement their products and services providing human analysis and interpretation where suitable for the consumer and profitable for the journalistic business. The traditional human brand based journalism which the NYT, the Washington Post, El País, La Nación,… among others represent, are merging at a steady pace with the DBMJ and an all digital and data based journalistic value chain. The old and new editorial functions are progressively being allocated to new working profiles and roles situated at renewed and specific value chain loci where the intersection of the human and the machine is being substituted by primarily artificial intelligence processes given its competitive advantages.

The obligations of automation, artificial intelligence, machine learning journalism should keep in place the norms, ethics and values transcribed into the new "all digital" software-hardware journalism. The relationship with the audience either performed by an all automatic "robotic news provider" with the supervision and mediation of human intelligence should have the obligation of preserving consumer rights and effective implementation of privacy and data management policies.

We provide in our research with the prototype of one of the latest predictive instruments and methods for research on algorithms, machine learning, automation, and news, the "distributed journalistic oriented sentiment analysis" tool for real time tweets DJOSA-tool, one of many tools for the DBMJ.

A new paradigm for contemporary empirical research, and rigorous conceptual development on digital journalism and data analysis has to be implemented in the coming years in the new data scenario for news production and distribution. Research orientated to empirical analysis with rich conceptual discussing will be presented in our case study. The promised land for journalisms which algorithms and automation is building will provide personalisation of content, faster news provision associated to preference curves. Apps are bound to be the utensil in engaging users in a mixed publicity funded and direct pay business model. Social networks and content providers will necessarily merge and search for quality content in order to retain consumers within their personalised environments.

In this article-research we describe and evaluate the application of Predictive Sentiment Analysis -PSA-, to a political communication case study through a real-time classifier of political opinions in Spanish tweets using machine learning methods and techniques both on a local computer and using distributed computing for Big Data problems. We present the pilot application and the first results of the designed data experiment and prototype. We describe the associated emerging methodologies and techniques and analyze the threats and opportunities that these innovations represent for political communication and other communicational research areas of interest. This prototype freely accessible-open sourced enabling the communication researcher to autonomously interpret the data given minimal prior training on the techniques and methods. It provides a scientific instrument designed for the understanding of media flows and political though and opinion.

The data paradigm has arrived as an unquestionable source of information concept for the studies of digital culture and digital media, communication and technology. Algorithms and data are today fundamental in order to effectively contrast former unexplorable hypothesis. It may be disruptive at the beginning but it is certainly a change of paradigm, from obscurity to potentially large quantities of data analysis and machine learning methodologies. The focus does not change, it is understanding communicational processes, but the instruments and data do change how we understand, study and research in our disciplines. The shift of focus on algorithms and data is positively disruptive for the ways in which we see our research and disciplines. It may even appear to limit the theoretical and methodological tools through which we increasingly try to understand mediation, the formation of identity, social life, politics and the creative industries.

There is a need to reformulate the theoretical and empirical perspectives and even paradigms on Communication Research and its relation with data acquisition, curation and interpretation. If we are to augment and diversify our perspectives at the Communication Research Academia, algorithms, machine learning and artificial intelligence are certainly must methods and instrument in bringing light to the still scientifically unexplored digital citizen-consumer. This research shows a methodological and instrumental methods to address communicational digital processes, from a complementary approach to existing scientific methods.

## References


Anderson, C.W. (2017) Social survey reportage: Context, narrative, and information visualization in early 20th century American journalism. Journalism 18:1, pages 81-100.

Ausserhofer, J., Robert Gutounig, Michael Oppermann, Sarah Matiasek, Eva Goldgruber. (2017) The datafication of data journalism scholarship: Focal points, methods, and research propositions for the investigation of data-intensive newswork. Journalism 17.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, 450- 453.

Carlson, M. (2015). The Robotic Reporter. Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. Digital Journalism, 3:3, pp. 416-431. http://dx.doi.org/10.1080/21670811.2014.976412

Cobb, W. N. W. (2015). Trending now: using big data to examine public opinion of space policy. Space Policy, 32, 11-16. Dörr, K.N. (2016) Mapping the field of Algorithmic Journalism. Digital Journalism, 4:6, pp. 700-722.

Dörr, K.N., Hollnbuchner, K. (2017) Ethical Challenges of Algorithmic Journalism, Digital Journalism, 5:4, 404-419, DOI: 10.1080/21670811.2016.1167612

Davies, K., Trevor Cullen. (2016) Data Journalism Classes in Australian Universities: Educators Describe Progress to Date. Asia Pacific Media Educator 26:2, pages 132-147.

Farida Vis (2013) Twitter as a reporting tool for breaking news. Digital Journalism, 1:1, 27-47, DOI: 10.1080/21670811.2012.741316 Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, 12. Groot-Kormelink, T. , Costera-Meijer, Irene (2014) Tailor-Made News, meeting the demands of news users on mobile and social media. Journal of Journalism Studies, Volume 15:5: Future of journalism: in an age of digital media and economic uncertainty. Pp. 632-641. http://dx.doi.org/10.1080/1461670X.2014.894367

Jung, J., Haeyeop Song, Youngju Kim, Hyunsuk Im, Sewook Oh. (2017) Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. Computers in Human Behavior 71, pages 291-298.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

Leetaru, K. (2012). Data mining methods for the content analyst: An introduction to the computational analysis of content. Routledge. Mark Coddington (2015) Clarifying Journalism's Quantitative Turn, Digital Journalism, 3:3, 331-

348, DOI: 10.1080/21670811.2014.976400 Marchetti, R., Ceccobelli, D. (2016) Twitter and Television in a Hybrid Media System. Journalism Practice 10:5, pages 626-644.
Momin M. Malik, Jürgen Pfeffer. (2016) A Macroscopic Analysis of News Content in Twitter. Digital Journalism 4:8, pages 955-979. Nicholas Diakopoulos (2015) Algorithmic Accountability, Digital Journalism, 3:3, 398-415, DOI: 10.1080/21670811.2014.976411
O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11(122-129), 1-2.
Spark Kafka Integration (2016). Spark Streaming + Kafka Integration Guide. Available at: http://spark.apache.org/docs/latest/streaming- kafka- integration.html
Seth C. Lewis (2015) Journalism In An Era Of Big Data, Digital Journalism, 3:3, 321-330, DOI: 10.1080/21670811.2014.976399
Shahin, S. (2016) When Scale Meets Depth: Integrating Natural Language Processing and Textual Analysis for Studying Digital Corpora. Communication Methods and Measures 10:1, pages 28-50.
Sormanen, N., Jukka Rohila, Epp Lauk, Turo Uskali, Jukka Jouhki, Maija Penttinen. (2016) Chances and Challenges of Computational Data Gathering and Analysis. Digital Journalism 4:1, pages 55-74.
Swart, J., Peters, C., Broersma, M. (2016) Navigating Cross-Media News Use. Journalism Studies 0:0, pages 1-20.
Swart, J., Chris Peters, Marcel Broersma. (2017) Repositioning news and public connection in everyday life: a user-oriented perspective on inclusiveness, engagement, relevance, and constructiveness. Media, Culture & Society, pages 016344371667903.
Tuomo Hiippala (2017) The Multimodality of Digital Longform Journalism, Digital Journalism, 5:4, 420-442, DOI: 10.1080/21670811.2016.1169197
Turck, M. & Hao, J. (2016). The Chart of the Big Data Landscape 2016 (Version 3.0). Available at: http://mattturck.com/big-data- landscape- 2016-v18-final/
Thurman, N. , Dörr, K.N, Kunert, J. (2017) When reporters get hand-on with robo-writing: Professionals Condier Automated Journalism`s Capabiities and Consequences, Digital Jounalism. http://dx.doi.org/10.1080/21670811.2017.1289819
Thurman, N. J., Schifferes, S., Fletcher, R., Newman, N., Hunt, S. & Schapals, A. K. (2016). Giving computers a nose for news: exploring the limits of story detection and verification. Digital Journalism, 4(7), pp. 838-848. doi: 10.1080/21670811.2016.1149436
van Zoonen, W., & Toni, G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. Computers in Human Behavior, 63, 132-141.
Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal of Advanced Research in Computer Science and Software Engineering, 2(6),282-292.
Wiesenberg, M., Zerfass, A., Moreno A. (2017) Big Data and Automation in Strategic Communication. International Journal of Strategic Communication 11:2, pages 95-114.
Wilson, T., Wiebe, J. & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.
Yeon-Lee, N., Kim, Y., Sang, Y. (2017) How do journalists leverage Twitter? Expressive and consumptive use of Twitter. The Social Science Journal 54:2, pages 139-147.

# Style, Singularity, and Substance: What Picture Editors Want from A.I.

Martin Schön
LMU Munich
martingeorg.schoen@gmail.com

Professor Neil Thurman
LMU Munich
neil.thurman@ifkw.lmu.de

**Abstract:** This paper aims to assess the opportunities and challenges inherent in the use of artificial intelligence in journalistic picture editing. To do this we built a tool that suggests images to illustrate news articles using keyword extraction and evaluated it using a qualitative online survey of professional picture editors and a simple, manual "suitability" heuristic. Our results show that the tool is able to return a factually "suitable" image about half the time, performing better on national or international stories than on those with a local or regional focus. However, the survey of picture editors revealed that whether an image matches a story's topic is not the only criteria used in image selection. Also important is whether the image: works within the space allocated to it on the page, matches the target publication's house style, has particular aesthetic qualities, and is original—helping the story and its respective publication to stand out from the competition. The development and deployment of artificial intelligence tools in journalistic picture editing will need to consider these contextual and artistic issues, as well as the resistance to automation that some professional picture editors expressed to us.

**Keywords:** Artificial intelligence, Image Selection, Keyword extraction, Machine learning, Picture editing

## Introduction

In August 2018, Getty Images launched 'Panels', an "artificial intelligence tool ... that recommends ... visual content to accompany a news article." The promise was that it could help picture editors create "better stories, more quickly" (Getty 2018). The launch is part of a trend for artificial intelligence to be applied to both editorial functions within news, as well as to the creation of still and moving images more widely. In this paper we describe a system that we have built that automatically selects images for news articles and evaluate that system with the help of professional picture editors in order that we can assess the opportunities and challenges inherent in the use of artificial intelligence in journalistic picture editing.

## Methods

We built a system that suggests images to illustrate news articles using keyword extraction. The tool takes the plain text of a news article as input, returning a search string that is used to query an image database. The tool ranks all terms that occur in the article according to three criteria. Firstly, term frequency: the number of times a term occurs. Secondly, first occurrence: the position of the term's first mention. Thirdly, entity category: a semantic categorisation of the term retrieved from the Thomson Reuters Open Calais tagging service. Examples of categories include HumanProtagonist and Location. The ranking is performed by a feedforward neural network.

The network is trained to classify good and bad search terms with a set of 100,000 terms generated from our own corpus of 20,000 BBC News articles. The highest ranked terms are compiled into a search query. In addition to this machine learning approach, a second ranking mechanism was developed. This statistical approach calculates the ranking score for each term directly from the term frequency and first occurrence values, without any prior learning involved.
A demonstration of the tool using the Getty Images API is available online (Schön 2018).

In order to evaluate the system we used a qualitative online survey of professional picture editors and a simple, manual "suitability" heuristic. The survey used a convenience sample (N=25) and asked picture editors about their work routines, with a focus on identifying tasks that had the potential to be automated and on how they selected images to illustrate articles. The editors were then given an opportunity to use our system by inputting text stories

to receive, automatically, suggestions for illustrative images. Following the interactive demonstration, the respondents were asked to discuss the system's strengths and limitations, and AI's future potential in their work.

## Findings

Our first evaluation used a simple, manual heuristic to determine the system's performance in terms of the general suitability of the images suggested for a given news story. We tested implementations of both ranking mechanisms—four different neural networks and one statistical approach—with 100 articles from the BBC News corpus. The Getty Images API was used as the image database. The resulting images were manually classified as either "suitable" or "unsuitable" with regard to the respective article. An image was deemed "suitable" if it illustrated the main topic of the story, but no judgement was made on the image's aesthetic qualities. Because all images returned came from the Getty Images database, they met Getty's minimum standards for sharpness, exposure, composition and so forth. The evaluation results show that both ranking mechanisms perform similarly well, with the neural networks performing slightly better. The tool worked better on articles without a local or regional focus, such as news about international politics, technology, science or business. On articles with a local focus, the statistical approach outperformed the networks by far. First occurrence proved to be the most powerful criterion for judging the suitability of a term as search query. On average around half the images returned were classified as "suitable."

Our second evaluation involved a qualitative online survey of professional picture editors. Most of the editors selected images to illustrate specific articles at least daily, but very few were aware of, or had used, any software that could help automate their routine tasks. Those who had did not come away very impressed. One had used software that could automatically crop images but found it "restrictive". Another had used the Getty 'Panels' product mentioned in the introduction but thought it was deficient, not being "smart or subtle enough."

The picture editors' feedback on our system was, on balance, more negative than positive. On the positive side some acknowledged that it was capable of suggesting suitable illustrative images quickly. One said they might use it "if I had a rush on or was stuck for ideas". Others suggested it might be useful for teams who did not have a designated, or experienced, picture researcher. On the negative side, a common criticism concerned the lack of relevance of some of the suggested images—a limitation our "suitability" heuristic also highlighted. The editors' survey revealed other shortcomings too. Some editors wanted the system to be able to suggest images from a wider variety of sources (not just Getty) and to show the costs of licencing particular images—not insurmountable technical challenges. However, some of the editors' other wishes present more of a challenge. Firstly, several mentioned that images do not only have to illustrate the content of a particular story but they need to be in keeping with the house style of the target publication. Secondly, an image has to be suitable for the space available for it on the page, which might preclude images with certain compositions or aspect ratios. Thirdly, editors emphasized the importance of images' aesthetics, for example their "beauty" and "visual impact." Finally, the importance of having "original" or "unique" images was emphasized, a thought encapsulated in this response from one of our respondents: "I want something different to what everyone else has."

## Conclusions

In this study we have highlighted some of the work being undertaken to apply artificial intelligence to the task of selecting images to illustrate stories. Our own system was described, and we demonstrated, both quantitatively and qualitatively, that it is capable of returning a factually "suitable" image about half the time. With further development this proportion should improve. However, our survey of professional picture editors revealed that, in the real-world, it is not enough for an image to match the topic of a story. For AI image selection tools—like ours and Getty's 'Panels'—to be useful they need to do more.

Two interesting challenges are, firstly, to make the tools contextually aware, both at the page level (where the image will appear) and at the outlet level (the publication's house style) and, secondly, to offer selections that are able to fulfil editors' desire to have "original" images that are "different" to their competitors. Two other requirements present a much greater level of challenge. Firstly, to select images that have the right aesthetic and

emotional qualities and, secondly, to build systems that do not make picture editors feel that the “creative” element of picture editing is being “taken away” from them.

## References

Getty (2018) “Getty Images Launches AI Tool to Transform Search for Media Publishers” 2 August, http://press.gettyimages.com/getty-images- launches-ai-tool-to-transform-search-for-media-publishers/
Schön, Martin (2018) Picpic Explorer. http://picpic-explorer.argonn.me/#/. Accessed 24 October 2018.

# Can data journalism really stimulate local news? A case study with media in the country side of Portugal

Ricardo Morais
University of Beira Interior / Labcom.IFP
ricardo.morais@labcom.ubi.pt

Pedro Jerónimo
University of Beira Interior / Labcom.IFP & CECS
pedrojeronimo.phd@gmail.com

**Abstract:** This study aims at outlining the development of data journalism at twenty-five Portuguese local media. The results presented are based on a survey to 107 journalists from local newspapers and radios. A content analysis evidencing data journalism published on their websites is also presented here, and confronts results obtained through the survey. The main findings suggest that despite the knowledge journalists have, there is no investment in data journalism because this type of practice is not considered to be a determining factor in attracting new audiences and approaching the most-assessed news. Considering the carried-out analysis, we propose, within the scope of the project in which we developed this study, some practices that local media can adopt that can help journalists to realize the potential of data journalism, but above all, to encourage them to adopt their practices.

**Keywords:** Data journalism; local news; Portugal; Remedia.Lab

## Introduction

In recent years much has been said about the possibilities of data journalism and how it can improve the journalistic field. But the truth is, as Rogers reveals, "Data journalism is not new", actually, the first example of data journalism dates back to 1821, "in the very first Guardian" and is related to a "list of schools in Manchester and Salford, with how many pupils attended each one and average annual spending" (Rogers, 2013, p. 60).

Therefore, this well-known form of journalism has just awakened in recent years because societies become increasingly digital and the amount of information available on networks grows. It was precisely this increase in the amount of information available that made data journalism become determinant in two levels: "1) analysis to bring sense and structure out of the never-ending flow of data and 2) presentation to get what's important and relevant into the consumer's head" (Meyer apud Gray, Chambers and Bounegru, 2012, p. 6). At a time where information is everywhere, the most important tasks are no longer search and gather, but filtering and verification. The role of journalists is, in this context, particularly significant, since they have the power to make sense of information. As Pilhofer says, data journalism "can include everything from traditional computer-assisted reporting (using data as a "source") to the most cutting-edge data visualization and news applications", but the ultimate goal remains the same: "providing information and analysis to help inform us all about important issues of the day" (Pilhofer apud Gray, Chambers and Bounegru, 2012, p. 6). In spite of that, there are some limitations in data journalism quality even in major media companies (Young, Hermida and Fulda, 2018).

Important for the production and dissemination of information in the new digital ecosystem, data journalism "may be the most powerful forum of collective journalistic sensemaking" (Anderson, 2019). This practice assumes particular importance in certain contexts, such as the local one. As Kristen Muller, a chief content officer at KPCC, says, "if local newsrooms are going to achieve digital sustainability, we must try new things. We need to experiment with different approaches to coverage and revenue" (2018). Therefore, proximity, that is a characteristic of local and regional media, can have in data journalism a unique possibility to "finding unique stories (not from news wires), and executing the watchdog function". Jerry Vermanen believes data journalism is crucial to regional newspapers, "because local newspapers have this direct impact in their neighborhood and sources become digitalized, a journalist must know how to find, analyze and visualize a story from data (Vermanen

apud Gray, Chambers and Bounegru, 2012, p. 7). Also Stefan Back (2018) pointed out that the engagement of journalists and civic technologists can be challenging to public service at a local level.

The question then arises as to whether these local media, where working conditions are often scarce since the number of journalists is low, understand the potential of data journalism and are ready to it. Sometimes they do and also have people in the newsrooms with that kind of knowledge (editorial and technical staff) but, unfortunately, that is not embedded in a company or editorial strategies (Jerónimo, 2015). Do local journalists see data journalism as a way to scrutinize the world and hold the powers accountable? Are journalists aware of data journalism techniques? Can they understand basic skills from traditional journalism just aren't enough in a digital era?
These are some of the questions that we seek to answer in this paper, through an analysis of a set of local media in the central region of Portugal, a significant part of the Portuguese media landscape and where the number of media has declined in recent years, due to the lack of public support and low audiences.

## Methods

In terms of research methods, we opted for the strategy of the study case, since it seems to us as a more adapted tool for the reality that we intend to study. For Yin (1989) case study is empirical research which consist in the analysis of a particular phenomenon in the real world, through different ways of collecting data. Rossman and Rallis (2003) consider that the case studies "seek to understand the larger phenomenon through a close examination of a specific case and therefore focus on the particular" (p. 104). Our case study is, in fact, a multiple case study (Yin, 1989, p. 52), which was characterized by the fact of performing in different local media at the same time. To answer our questions, we collect data through a survey with journalists from twenty-five local newspapers and radios, but we conduct also a content analysis in search of data journalism examples, published on the websites of the media investigated. We act according to the proposals of Yin (1989), who advocate the use of different data sources, i.e., "multiple sources of evidence" (p. 23). The journalistic projects chosen in this study aim at representing the central area of Portugal, one of the most affected at the level of media communication foreclosure. On the other hand, these media are also part of the project Remedia.Lab in which we try "to diagnosis the current situation of local/regional media, promoting experimental tools and strategies to strengthen their business model, increasing their innovation degree and improving their connection with the public".

## Findings and Argument

Results from the survey show that journalists see as very important having knowledge in web-scrapping, obtaining tools for analysis and data collection, as well as gaining knowledge in the creation of infographics and data presentation. Furthermore, answers collected from the local journalists surprisingly show that these professionals have a good knowledge in a set of skills, such as web-scraping, data visualization and presentation. However, findings show that journalists do not consider the use of data journalism as the most effective approach to attract new audiences as well as to retain them. Also the lack of company and editorial strategies can help to discourage this way of thinking (Jerónimo, 2015). This is the kind of data that we seek to explore in this work, especially since several studies indicate that data journalism can help to revitalize local journalism in general, and more particularly, small local journalism offices. Can these results help explain the limited number of journalism works based on data journalism we found on media websites? If, as the survey shows, knowledge in data journalism techniques exist, is the lack of investment due to insufficiency of human and material resources? These are some of the results that we will explore in this work trying to confront this reality with the amount of information that is now available on the network, but also questioning the accessibility of the data for journalists' work.

## Conclusions

Although our findings show that local media journalists have technical knowledge to manage data journalism, those skills are not invested in news production. Such situation can be explained by the lack of strategies, the presence of small newsrooms and a traditional news making culture, as it is especially evident in newspapers. On the other hand, we cannot ignore findings of previous studies that identify "key actors" in the newsrooms:

journalists with the ability to innovate in their field, on their own, even without recurring to existing strategies (Jerónimo, 2015).

With this work, we would like to point out that the lack of data journalism in Portuguese local media is an opportunity for journalistic projects to assume that there are opportunities in the journalistic field that must be explored to capture and maintain audiences. Identifying, encouraging and helping the key players in the essays are some of the steps that need to be taken next. These results, together with the evaluation of the type of data that public services and governments provide, constitute important knowledge that can be transmitted in the form of advice to the local media, in order to help them implement data journalism in their essays. These results will also be important for public and private entities at local and regional level to openly disclose their data.

## References

Alexandre, I. A. R. (2014). Jornalismo de Dados: o estado da arte nos jornais generalistas diários em Portugal. Mestrado em Novos Media e Práticas Web. Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa. Available at https://run.unl.pt/handle/10362/13615 [Accessed April 3, 2019].
Anderson, C.W. (2019). Genealogies of Data Journalism. In: J. Gray and L. Bounegru, eds., The Data Journalism Handbook 2: Towards a Critical Data Practice. European Journalism Centre and Google News Initiative. Available at https://datajournalismhandbook.org/handbook/two/situating-data-journalism/genealogies-of-data-journalism [Accessed April 3, 2019].
Gray, J. and Bounegru, L. (Eds.) (2019). The Data Journalism Handbook 2: Towards a Critical Data Practice. European Journalism Centre and Google News Initiative. Available at https://datajournalismhandbook.org/handbook/two#situating-data-journalism [Accessed April 3, 2019].
Baack, S. (2018). Practically Engaged, Digital Journalism, 6(6), 673-692, DOI: 10.1080/21670811.2017.1375382
Gray, J., Bounegru, L. and Chambers, L. (2012). The Data Journalism Handbook. How journalists can use data to improve the news. Califórnia:
O'Reilly Media. Available at http://datajournalismhandbook.org/ [Accessed April 3, 2019].
Jerónimo, P. (2015). Ciberjornalismo de proximidade: redações, jornalistas e notícias online. Covilhã: LabCom Books.
Martinho, A. I. P. (2013). Jornalismo de dados: contributo para uma caracterização do estado da arte em Portugal. Dissertação de Mestrado. ISCTE-IUL. Available at http://hdl.handle.net/10071/8329 [Accessed April 3, 2019].
Meyer, P. (2002). Precision Journalism: a reporter's introduction to social science methods. Maryland: Rowman & Littlefield Publishers. Rogers, S. (2013). Facts are Sacred: The Power of Data. London: Guardian Books.
Rossman, G. and Rallis, S. (2003). Learning in the Field: An introduction to qualitative research. Thousand Oaks (California): Sage Publications. Yin, R. (1989). Case Study Research – Design and Methods. London: Sage Publications.
Young, M. L., Hermida, A. and Fulda, J. (2018). What Makes for Great Data Journalism?, Journalism Practice, 12(1), 115-135. DOI: 10.1080/17512786.2016.1270171
Young, W. D. (2018). Data Journalism Goes Undercover. Available at https://www.niemanlab.org/2019/01/data-journalism-goes-undercover/ [Accessed April 3, 2019].