Big Data Analysis of News and Social Media Content

Ilias Flaounas, Saatviga Sudhahar, Thomas Lansdall-Welfare,
Elena Hensiger, Nello Cristianini (*)

Intelligent Systems Laboratory, University of Bristol

(*) corresponding author

Abstract

The analysis of media content has been central in social sciences, due to the key role that media plays in shaping public opinion. This kind of analysis typically relies on the preliminary coding of the text being examined, a step that involves reading and annotating it, and that limits the sizes of the corpora that can be analysed. The use of modern technologies from Artificial Intelligence allows researchers to automate the process of applying different codes in the same text. Computational technologies also enable the automation of data collection, preparation, management and visualisation. This provides opportunities for performing massive scale investigations, real time monitoring, and system-level modelling of the global media system. The present article reviews the work performed by the Intelligent Systems Laboratory in Bristol University towards this direction. We describe how the analysis of Twitter content can reveal mood changes in entire populations, how the political relations among US leaders can be extracted from large corpora, how we can determine what news people really want to read, how gender-bias and writing-style in articles change among different outlets, and what EU news outlets can tell us about cultural similarities in Europe. Most importantly, this survey aims to demonstrate some of the steps that can be automated, allowing researchers to access macroscopic patterns that would be otherwise out of reach.

Introduction

The ready availability of masses of data and the means to exploit them is changing the way we do science in many domains (Cristianini, 2010; Halevy et al., 2009). Molecular biology, astronomy and chemistry have already been transformed by the data revolution, undergoing what some consider an actual paradigm shift. In other domains, like social sciences (Lazer et al., 2009; Michel et al., 2011) and humanities (Moretti, 2011), data-driven approaches are just beginning to be deployed. This delay was caused by the complexity of the social interactions and the unavailability of digital data (Watts, 2007). Examples of works that deploy computational methods in social sciences include: the study of human interactions using data from mobile phones (Onnela et al., 2007; González et al., 2008); the study of human interactions in online games and environments (Szell et al., 2010); or automating text annotation for political science research (Cardie et al., 2008).

In particular, the analysis of media content in the social sciences, has long been an important concern, due to the important role that media play in reflecting and shaping public opinion (Lewis, 2001). This includes the analysis of both traditional news outlets, such as newspapers and broadcast media, and modern social media such as Twitter where content is generated by users. The analysis of news content is traditionally based on the coding of data by human coders. This is a labour-intensive process that limits the scope and potential of this kind of investigation. Computational methods from text mining and pattern recognition, originally developed for applications like e-commerce or information retrieval, can be deployed to automate many aspects of data collection, annotation, analysis and visualisation.

Our group at the Intelligent Systems Laboratory in Bristol has worked for the past six years in this direction, creating technologies to support Social Sciences investigations and to demonstrate their use on real-world large scale datasets. We have built a computational infrastructure which is capable of gathering, storing, translating, annotating and visualising vast amounts of news items and social media content (Flaounas et al., 2011). Currently our system tracks 1,100 news outlets in 22 languages, and tweets from 54 UK locations. Overall the system has generated over one Terabyte (equivalent to 1000 Gigabytes) of data over the past four years.

This article reviews part of our research in media content analysis, which we call the MediaPatterns project[1]. The technical details have been avoided in an attempt to emphasise the capabilities of data-driven approaches in social sciences research. For each study we provide references to the original articles and websites where the results were first presented. It is important to mention that each study was performed by using different methodologies depending on the task. In the present article we will review the following: Sentiment analysis of Twitter content (using keyword matching and time-series analysis techniques); Narrative analysis of US Elections (using natural language processing and network analysis techniques); Comparison of news outlets based on topics covered, writing style, and gender bias (using machine learning and natural language processing techniques); Modelling of reader preferences (using machine learning techniques); Network analysis of the EU mediasphere (using clustering and network analysis techniques); and event detection in textual streams (using statistical approaches).

Sentiment Analysis of Twitter Content

Measuring the current public mood is a challenging task. The traditional approach would require questioning a large number of people about their feelings. Social media, such as Twitter or Facebook, can easily become a valuable source of information about the public due to the fact that people use them to express their feelings in public.

As demonstrated in our study by Lansdall-Welfare et al. (2012) it is feasible to capture the public mood by monitoring the stream of Twitter data. The dataset that was analysed was comprised of 484 million tweets that were generated by more than 9.8 million users, between July 2009 and January 2012. The data were collected from the 54 largest cities in the UK. We focused on tracking four moods which are “Fear”, “Joy”, “Anger” and “Sadness”. For each mood, we track a long list of associated words and we count the frequencies that these words appear in tweets. This process generates one timeline of the volume of related tweets for each emotion. The further analysis of these timelines reveals that each of the four emotions changes over time in a rather predictable manner. To give an example, we found a periodic peak of joy around Christmas and a periodic peak of fear around Halloween. More surprisingly, we found that negative moods started to dominate the Twitter content after the announcement of massive cuts in public spending on October 2010. Also there was a significant increase in anger in the weeks before the summer riots of August 2011. In Fig. 1 we plot the mood levels for the period of study and we visualise them as a facial expression using the Grimace tool. In the project website we have an

[1] MediaPatterns website: http://mediapatterns.enm.bris.ac.uk

interactive demo that can be used to explore our data[2].

A detailed interpretation of the results of this study is a challenging task. A scholar needs to understand the causes of the measured statistical quantities and associate them with specific events that affected society. The interpretation of the results can not be automated by using computational methods and it is a task that needs to be undertaken by social scientists.

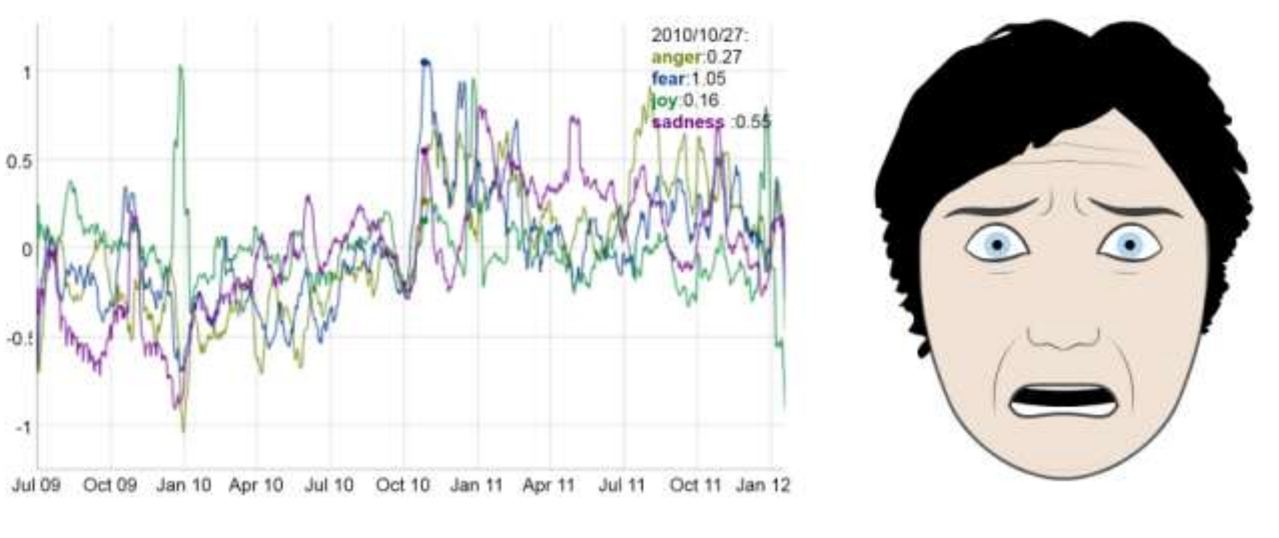

Figure 1. Analysis of 484 million tweets over a period of three years can reveal changes in public mood. We track four basic emotions and we can associate mood changes with important events (Lansdall-Welfare et al., 2012).

Narrative Analysis of US Elections

Content analysis sometimes involves the identification of basic narrative information in a corpus. For example, determining the key actors, the key actions, and how actors interact. Quantitative Narrative Analysis (Franzosi, 1987; Earl et al., 2004) involves the process of extracting Subject-Verb-Object (SVO) triplets and their study in order to analyse a corpus, while “distant reading” in humanities transforms novels into networks as a first step of analysis (Moretti, 2011). The extraction of SVO triplets and turning them into a network is a process that can be automated and scaled up, enabling the analysis of big corpora.

In our work by Sudhahar et al. (2012), we used computational methods to perform narrative analysis of the US Elections 2012. The dataset we analysed is comprised of 125,254 articles that were automatically categorised as being related to the US Elections using the machine learning Support Vector Machines (SVM) approach (Cristianini, 2000). We pre-processed the corpus text using GATE (Cunningham, 2002) tools and then we used Minipar parser (Lin, 1998) to extract Subject-Verb-Object triplets. We identified 31,476 actors from these triplets. A statistical analysis identified the key actors and actions. Furthermore, we separate actions into two categories, endorsement and opposition. Finally, the key actors and actions were used to infer the network of the US Elections. This network is presented in Fig. 2 were the protagonists of the Elections, “Obama” and “Romney”, are easily observed to dominate the network.

[2] Mood Changes in Twitter Content: http://mediapatterns.enm.bris.ac.uk/mood/

Exploring this network can give a useful insight on the topics that are most discussed during the elections, who the most dominant or least reported people are, how often the key actors are presented by media as endorsing or opposing topics etc. Our results can be accessed online through an interactive website that we built for the monitoring of the elections[3]. The website is updated daily for the duration of the elections. A further analysis would require a social scientist to find and deeply understand the causes and associate them with raw data in order to interpret the patterns that emerge.

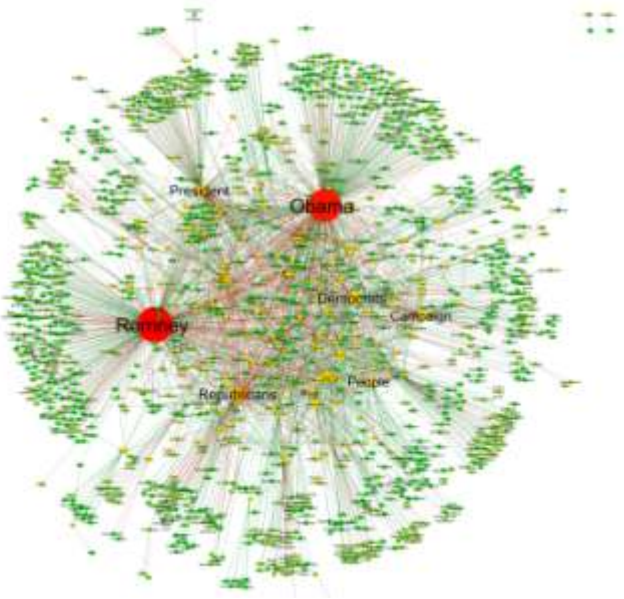

Figure 2. The key actors and their relations automatically extracted from a corpus of 125,254 articles related to US Elections 2012 (Saatviga et al. 2012).

Comparison of News Outlets on Topics, Writing Style and Gender Bias

Detection of differences or biases among news outlets is a popular field of research in media studies. Nowadays, several aspects of content analysis can be automated: using machine learning techniques such as Support Vector Machines (Cristianini et al., 2000) we can analyse text and detect general topics, such as Politics or Sports, that are present in a corpus; we can use natural language processing techniques to quantify properties of the writing style of the text or detect mentions of people and their gender.

In our work by Flaounas et al. (2012), we analysed a large dataset comprised of 2.5 million news articles collected from 498 different news outlets over a period of 10 months. For each article we identified the general topics that it covered, the people that they were mentioned and their gender (Cunningham et al., 2002), as well as two basic writing style properties, namely the readability and linguistic subjectivity. Readability is a measure of how easy it is to read a text calculated by taking into account the length of sentences and the length of the words in syllables (Flesch, 1948). Linguistic subjectivity is a measure of the usage of words that carry sentiment, i.e. the more words with sentiment present in an article, the more linguistically subjective it is (Pang, 2008).

The computation of the aforementioned quantities allows different comparisons of sets of articles. For example, for the articles of each topic we calculated: a) the average readability - finding that articles about sports are the easiest to read while articles on

[3] ElectionWatch: http://electionwatch.enm.bris.ac.uk

politics are the hardest to read; b) Linguistic subjectivity - finding that articles about fashion and arts are the most subjective while business articles were the most objective; c) the ratio of males over females – finding that articles about fashion and art mention the most women while articles about sports are male dominated. Furthermore, we directly compared 15 major US and UK newspapers on which topics they tend cover more often; their writing style and also the ratio of males over females that their articles mention. In Fig. 3 we visualise the comparison of outlets based on their writing style: outlets with similar writing style are closer together.

The approach for automatically detecting differences between outlets, people, topics, countries etc. is straightforward. The combination of large amounts of data in digital format with the application of machine learning and natural languages techniques can help social scientists answer a diverse range of research questions that could not be posed before.

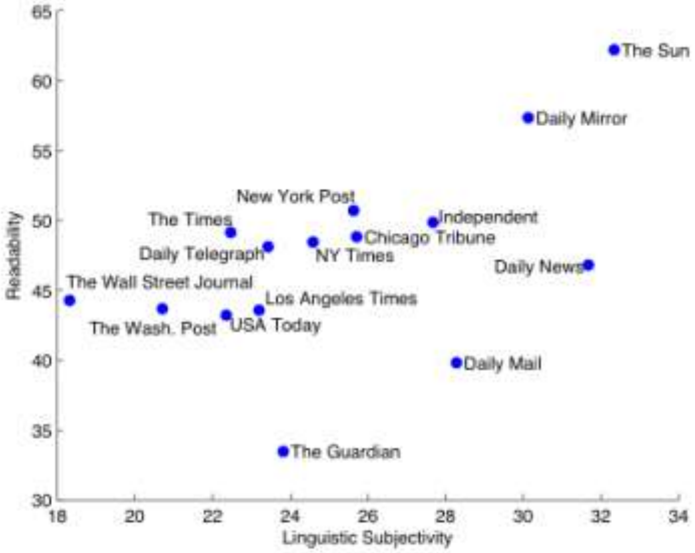

Figure 3. Newspapers compared based on their Readability and Linguistic Subjectivity. The illustration is based on the analysis of a dataset comprised of articles from US and UK newspapers (Flaounas et al. 2012).

The EU News Media Network

A recent technology that enables new avenues of research is machine translation. The technology is not mature enough to create human quality translations but nevertheless allows a scholar to access textual material written in most languages. Furthermore, machine translation can be combined with other computational methods in order to analyse texts written in non-English languages using computational tools developed for English. This reduces the costs of developing specialised tools for each language separately.

In our paper by Flaounas et al. (2010), we showed how machine translation allowed the content analysis of a large corpus comprised of 1.3 million articles published by main news media in the 27 EU countries over a period of six months. Those articles were written in 22 different European languages. All non-English articles were machine-translated into English before further processing. In this research, we used the machine learning approach of clustering to identify sets of articles that refer to the same event. Then statistical approaches were used to infer the network of the EU news media. In that network, two outlets are connected if they tend to publish articles about the same events. In Fig. 4 we illustrate a sparse version of that network. It can be observed that news outlets, represented as nodes of the network, are organised in communities. An interesting and expected result is that outlets from the same country belong to the same community, i.e. they are interested in the same stories.

A deeper analysis of the structure of the network revealed some less expected results. For example, that the EU media system reflects the geographic, economic, and cultural relations between the EU countries. Also, we showed that the deviation of the content of an EU news outlet from the "average" EU media content is significantly correlated with the year of accession to the EU of the country that the outlet is based (correlation 48.94%, *p*-value<0.01). Interestingly enough, the editors of 215 different news outlets from 27 countries made their independent choices over what stories they will cover; but these choices shape the EU media sphere in a way that reflects relations between EU countries.

The detection of these subtle signals in a statistically rigorous way required the analysis of a massive dataset and the deployment of methodologies such as machine translation, data clustering and network analysis. To conclude, we showed that ideally, the language differences should not be an obstacle for a scholar to formulate research questions and analyse data.

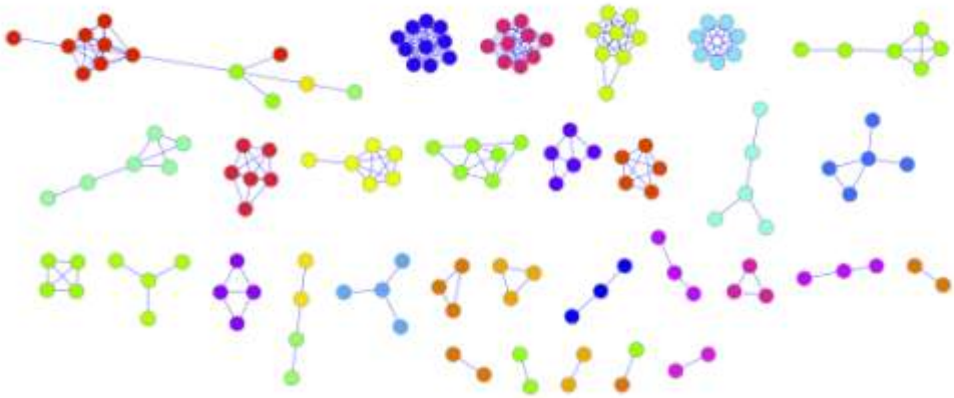

Figure 4. The communities of the network of EU news outlets. Nodes represent outlets and two outlets are connected if they tend to cover the same stories. Outlets from the same country are coloured with the same colour. To infer this network we used more than 1.3 million articles written in 22 different languages (Flaounas et al., 2010).

What Makes Us Click

Some online news outlets highlight their "Most Popular" stories. This is a valuable source of information since it can be used to model the reading preferences of their audience. The modelling starts by forming a large number of pairs of articles. Each pair is comprised of one popular article and one non-popular article, i.e. an article that was published on the main page of the outlet, on the same day, but it didn't become popular. Then a machine learning algorithm, namely Ranking-SVM, can be used to create a statistical model that captures the words that tend to appear in popular articles (Hensinger et al., 2012a). The same model can be used to predict if any article will become popular or not.

In the study by Hensinger et al. (2012b), we showed how we can model the reading preferences of the audiences of 14 news outlets. We collected popular and non-popular articles over a period of 20 months forming in total 2,432,148 pairs of articles. For each of the 14 news outlets we created a model of the preferences of their readers. These models had an average performance of 70.6% in predicting which article between a pair of articles has a better chance of becoming popular. In Fig. 5 we present as an example the word cloud of the model for “News.com.au” which shows the words that trigger the attention of the audience and the words that contribute less to the appeal of an article.

Similar approaches to the one we present can also be used to identify the preferences of people, by monitoring their online behaviour - simply checking what they click on. Both approaches offer a coarse understanding of people's behaviour. Of course, a deeper understanding would require a more traditional approach of questioning and interviewing people and reasoning about their answers.

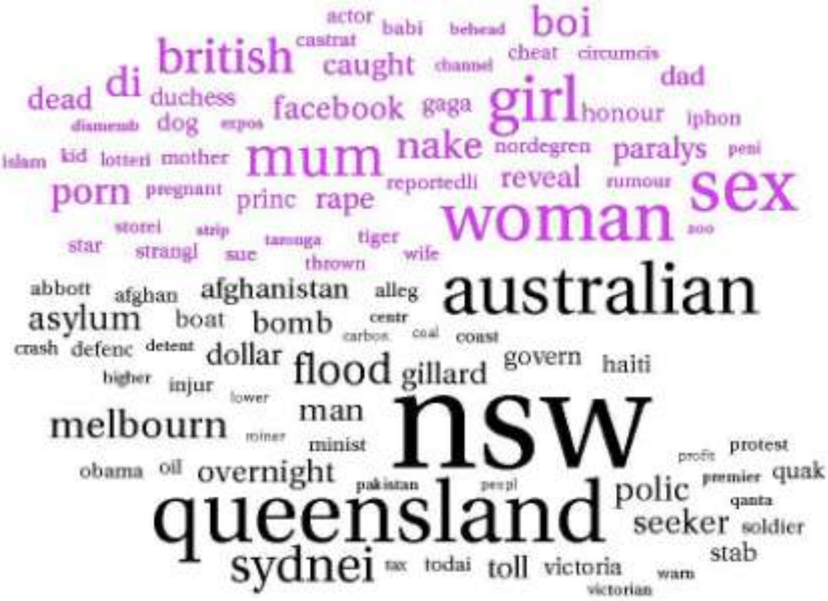

Figure 5. The word cloud presents in pink the words that are present in high-appeal articles published by “News.com.au” and in black the words that are present in low-appeal articles. The size of each word reflects the contribution of the word, either positive or negative, to appeal. The analysed dataset contained 2.4 million of pairs of popular and non-popular articles (Hensinger et al. 2012b).

Event Detection in Textual Streams

There are various ways in which big-data can be found “in the wild” and one increasingly important format is that of textual data streams, such as those generated by news feeds, social media or personal communications. When monitoring a data stream, an important issue is detecting change points. These are any variations in the statistical properties of its content that might signal important events.

In the research by Snowsill et al. (2010), we demonstrated one possible approach based on tracking sequences of words (*n*-grams) using a suffix tree methodology. We applied it to the New York Times corpus on articles published during a period of 10 years, from January 1996 until June 2007. The dataset we analysed contained in total 1,035,263 articles. Using a statistical approach we detected and ranked the major events. We plotted a timeline for each event for the whole period of the 10 years we analysed. In Fig. 6 we illustrate a selection of events. For example, detected events include “Princess Diana”, “Iraq”, “World Trade Center”, and “Afghanistan”. Note, that these major events were detected in a fully automated way and no-human supervision or interaction was necessary at any point during the whole process.

Once more the interpretation of these results, their deeper understanding and the extraction of useful knowledge out of them is a task for a social scientist. The computational approach can only highlight where the scholar should focus their attention.

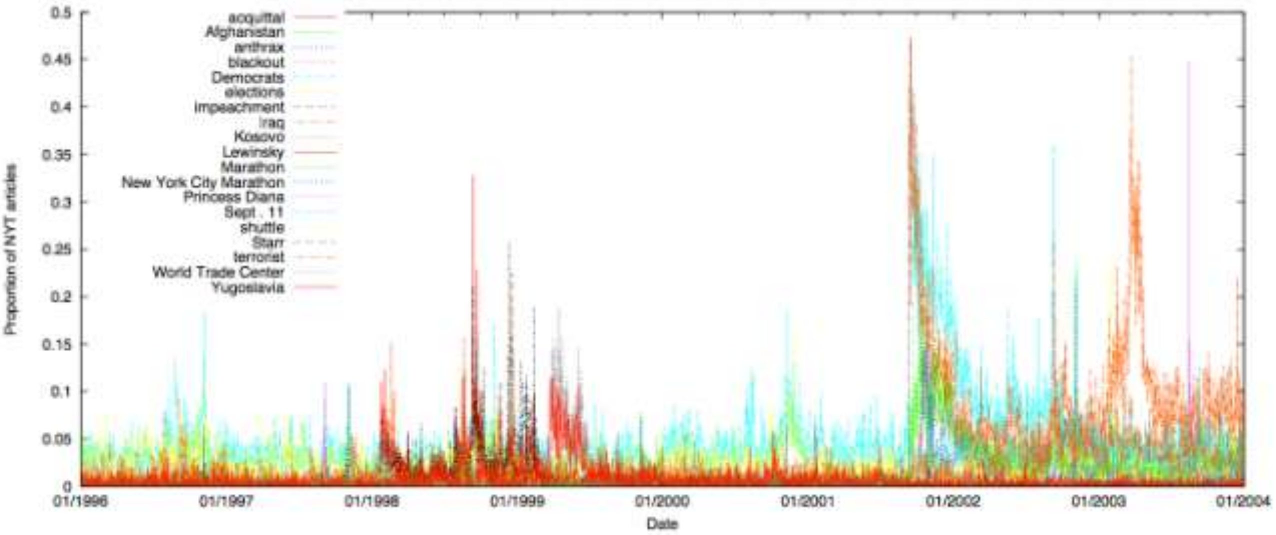

Figure 6. We analysed 1,035,263 articles from the New York Times corpus published over a period of ten years and we detected the most significant events reported (Snowsill et al., 2010).

Conclusions

The capability to gather vast amounts of digital data and the computational methodologies to analyse them are changing many aspects of science. The research questions that can be answered using data-driven approaches evolve and are multiplied based on the progress of pattern-analysis and data-mining technology. Machines provide the opportunities for extracting valuable information from big data and performing the repetitive and mundane tasks. Their capabilities to digest large amounts of data, far more than any researcher could analyse, are compensating for their inaccuracies and lack of reasoning. Of course, in many cases there is no substitute for human interpretation and judgement.

The boundary between tasks that can be solved by machines and those that can not is constantly changing. This is particularly true in the case of text analysis technologies. For example, just a decade ago machine translation could not be listed among the tools available to social scientists. The technologies of text-mining and text-analysis are evolving quickly, driven by strong commercial pressure, and as a side effect can lead to new tools for the social sciences. It is certainly worth keeping an eye on which new tools become available, as they could provide help in answering entirely new research questions. At the same time, entirely new types of data are constantly being released, ranging from news to social media content, from political proceedings to the content of millions of books (Michel et al., 2011). The opportunities for progress are real, when we properly combine these two types of resources.

The work surveyed in this paper demonstrates technologies that can reliably detect topics, sentiment, persons, events and relations among actors, applied on millions of documents. Computational methods can also retrieve them, translate them and code them, all on a scale that cannot be matched by human analysts. We followed common practices in text analysis, by calibrating the tools we deployed on reference datasets that have been annotated by humans, so that we know the precision of our automatic annotation. For example to calibrate topic detection algorithms we used the public datasets of New York Times (Sandhaus, 2008) and the Reuters (Lewis et al, 2004), while for discovering the gender of people we used the public database Freebase[4]. While this is clearly not perfect, there are applications for which this is sufficient. It is important to notice that data annotated by human researchers during their own studies can provide a very valuable resource for the training and evaluation of artificial tools, and should be shared and be open, much like is done in biology (Let data speak to data, 2005).

One important aspect of using big data in research has been mentioned only marginally in this survey: the need to visualise the data and the relations discovered in it, in order to help analysts make sense of the large scale datasets. We have demonstrated how political relations can be represented as networks, how mood can be represented as a face expression, but much more can be done in this direction.

While we have focused on the analysis of textual news content, and any other structure we derived from it (e.g. networks, maps and timelines), many researchers in the area of Computational Social Science (CSS) focus on different types of data: for example social networks, or search engine queries. In most of those cases, all data from CSS share the feature of being mostly unstructured, that is not naturally organised in rigorous structures, which is instead the norm in engineering. The exploitation of unstructured data is becoming a major area of concern in computer science and we can expect fast progress in the next few years.

As we move quickly towards a new way of doing science, we can see both opportunities and risks. It is important to keep both of them in mind, and to avoid unnecessary hype. There are many cases where forcing the culture of engineering on a domain such as the social sciences, but also the natural sciences, would be a mistake. We should keep in mind that these tools can only hope to complement – never to replace – the role of analysts, with their capability to understand context, to interpret and to incorporate their findings into a coherent narration. Data-driven technologies can be useful when they allow analysts to become aware of macroscopic trends that would be otherwise out of reach, but we should not delude ourselves that they have any chance of ever taking their place.

Acknowledgments

I. Flaounas and N. Cristianini are supported by European Community's Seventh Framework Programme under grant agreement N° 270327 (CompLACS Project). Authors would like to acknowledge the former members of the group, namely Omar Ali, Vasileios Lampos, Tristan Snowsill, and Marco Turchi for their contributions to the publications that were reviewed in this paper.

References

Let data speak to data (2005), (Editorial), Nature 438, pp. 531.

[4] Freebase: http://www.freebase.com

Cardie, C, Wilkerson, J (2008) “Text annotation for political science research”, Journal of Information Technology & Politics, 5(1), pp. 1–6.

Cristianini, N, Shawe-Taylor, J (2000) “An Introduction to Support Vector Machines and other Kernel-based learning methods”, Cambridge University Press.

Cristianini, N (2010), “Are we there yet?”, Neural Networks, Elsevier, 23(4), pp. 466-470.

Cunningham H, Maynard D, Bontcheva, K, And Tablan, V (2002) “GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications”, In Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics, pp. 168–175.

Earl J, Martin A, McCarthy J, Soule S (2004) The use of newspaper data in the study of collective action. Annual Review of Sociology 30:65–80.

Halevy, A, Norvig, P and Pereira, F (2009), “The unreasonable effectiveness of data”, Intelligent Systems, IEEE, vol. 24, pp. 8-12.

Flaounas, I, Turchi, M, Ali, O, Fyson, N, Bie, TD, Mosdell, N, Lewis, J & Cristianini, N (2010), 'The Structure of EU Mediasphere' PLoS ONE, vol 5 (12).

Flaounas, I, Ali, O, Turchi, M, Snowsill, T, Nicart, F, De Bie, T, Cristianini, N (2011) “NOAM: news outlets analysis and monitoring system”, SIGMOD Conference, ACM, pp. 1275-1278.

Flaounas, I, Ali, O, Lansdall-Welfare, T, De Bie, T, Mosdell, N, Lewis, J & Cristianini, N (2012), 'Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content: topics, style and gender', Digital Journalism, Routledge, vol 1, no. 1.

Flesch, R (1948) “A New Readability Yardstick”, *Journal of Applied Psychology* 32, pp. 221–233.

Franzosi R (2010) Quantitative Narrative Analysis. Sage Publications Inc, Quantitative Applications in the Social Sciences, pp. 162:200.

González, M, Hidalgo, C, and Barabási, A (2008), Understanding individual human mobility patterns. Nature, 453(7196), pp. 779–782.

Hensinger, E, Flaounas, I, Cristianini, N (2012a) “What makes us click? Modelling and Predicting the Appeal of News Articles”, Proc. of International Conference on Pattern Recognition Applications and Methods, pp. 41-50.

Hensinger, E, Flaounas, I, Cristianini, N (2012b) “The Appeal of Politics on Online Readers”, In: Internet, Politics, Policy 2012:Big Data, Big Challenges?, Oxford, Available at: http://microsites.oii.ox.ac.uk/ipp2012 [Accessed 1st Nov. 2012].

Lansdall-Welfare, T, Lampos, V & Cristianini, N (2012), 'Effects of the Recession on Public Mood in the UK'. In Mining Social Network Dynamics (MSND) session on Social Media Applications in News and Entertainment at WWW '12., ACM, pp. 1221 – 1226.

Lazer, D, Pentland, A, Adamic, L, Aral, S, Barabasi, A-L, Brewer, D, Christakis, N, et al. (2009) “Computational Social Science”, science 323, pp. 721-723.

Lewis, J (2001) “Constructing Public Opinion, New York: Columbia University Press.

Lewis, D, Yang, Y, Rose, T G, and Li, F (2004) “RCV1: A New Benchmark Collection for Text Categorization Research”, *Journal of Machine Learning Research* 5, pp. 361–397.

Lin D, (1998), “Dependency-Based Evaluation of Minipar”, Text, Speech and Language Technology 20 pp. 317–329.

Michel, J-B, Shen, Y K, Aiden, A P, Veres, A, Gray, M K., et al. (2011) “Quantitative

Analysis of Culture Using Millions of Digitized Books”, Science 331, pp. 176–182.

Moretti, F (2011) “Network theory, plot analysis”. New Left Review 68, pp. 80–102.

Onnela, J, Saramäki, J, Hyvönen, J, Szabó, G, Lazer, D, Kaski, K, Kertész, J, and Barabási, A (2007) “Structure and tie strengths in mobile communication networks”, Proceedings of the National Academy of Sciences, 104(18), pp. 7332–7336.

Pang, B, Lee, L (2008) “Opinion mining and sentiment analysis”, *Foundations and Trends in Information Retrieval* 2, pp. 1–135.

Sandhaus, E (2008) “The New York Times Annotated Corpus”, The New York Times Company, Research and Development.

Snowsill, T, Flaounas, I, De Bie, T, Cristianini, N (2010), 'Detecting events in a million New York Times articles'. In Proc. of Machine Learning and Knowledge Discovery in Databases, Springer, pp. 615-618.

Sudhahar, S, Lansdall-Welfare, T, Flaounas, I, Cristianini, N (2012), 'ElectionWatch: Detecting Patterns In News Coverage of US Elections'. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 82-86.

Szell, M, Lambiotte, R, Thurner, S (2010) “Multirelational organization of large-scale social networks in an online world”, Proceedings of the National Academy of Sciences, 107(31), pp. 13636–13641.

Watts, D (2011), “Everything is Obvious: Once You Know the Answer: How Common Sense Fails”, Crown Business, New York.

Watts, D (2007) “A twenty-first century science”, *Nature* 445, pp. 489.