On the Coverage of Science in the Media: A Big Data Study on the Impact of the Fukushima Disaster

Thomas Lansdall-Welfare*, Saatviga Sudhahar*, Giuseppe A. Veltri† and Nello Cristianini*
*Department of Computer Science, University of Bristol, Bristol, United Kingdom
†Department of Media and Communication, University of Leicester, Leicester, United Kingdom
Email: Thomas.Lansdall-Welfare@bris.ac.uk, Saatviga.Sudhahar@bris.ac.uk, Nello.Cristianini@bris.ac.uk, gv35@le.ac.uk

***Abstract*—The contents of English-language online-news over 5 years have been analyzed to explore the impact of the Fukushima disaster on the media coverage of nuclear power. This big data study, based on millions of news articles, involves the extraction of narrative networks, association networks, and sentiment time series. The key finding is that media attitude towards nuclear power has significantly changed in the wake of the Fukushima disaster, in terms of sentiment and in terms of framing, showing a long lasting effect that does not appear to recover before the end of the period covered by this study. In particular, we find that the media discourse has shifted from one of public debate about nuclear power as a viable option for energy supply needs to a re-emergence of the public views of nuclear power and the risks associated with it. The methodology used presents an opportunity to leverage big data for corpus analysis and opens up new possibilities in social scientific research.**

***Keywords*-Data analysis; Text mining; Knowledge discovery; Computational linguistics;**

I. Introduction

The portrayal of scientific and technological objects in the mass media plays a fundamental role in our understanding of how such issues are discussed and interpreted in the public sphere and constitute common sense knowledge [1], [2]. There is a wide scientific literature on this topic that combines communication theories and studies of mass media influences (e.g. agenda setting theory [3]; second level agenda setting [4]; cultivation theory [5]; agenda building [6]; second level agenda building [7]) with public understanding of science research. In particular, studying how mass media frame an emerging technology is important for observing definitions and associated meanings that are legitimized or stigmatized.

Nuclear power has been studied in the past by an extensive body of research with findings that have revealed the negative associations and imagery (e.g. accidents, destruction, contamination, mushroom clouds, child cancer etc.) often linked with such technology [8]–[10]; and which have been variously described by expressions like "nuclear fear" [11] and "nuclear stigma".

Studies in this area are focused on the idea that in the public sphere there are competing definitions, in what is a complex game played for the control of semantics in the public sphere [12]. Considering that definitions are not just technical issues, but are a matter of framing for the purpose of opinion and attitude formation and for regulation, competing representations in the media is a field where the battle "is being waged in the arena of language, as much as that of science" [13].

In this context, we analyzed the impact of a major event, the disaster at the nuclear power plant in Fukushima Daiichi (Japan, 11th March 2011) on media representations of nuclear power before and after. Media representations can exacerbate people's risk perception of nuclear power and a previous study has found such an effect in the US precisely in the case of the recent Japanese accident [14].

To this end, we use a combination of measurements before and after the event on a corpus of millions of news articles to detect the impact in the media and the changes in reporting of nuclear power thereafter. We focus on three particular aspects: the evolution of attention (salience) and sentiment of nuclear power, revealing the change in the overall volume of science articles covering the topic and the sentimental framing of the media coverage; networks of the actors and actions linked to nuclear power along with its action clouds, allowing us to detect the shift in actors involved in the public debate around nuclear power and also discover the new actions taking place in the debate; and finally the network of topics, universities and diseases associated with nuclear power, showing the changing latent representation of nuclear power as presented by online news media.

II. Data Description

We gathered over 5 million science articles between 1st May 2008 and 31st December 2013 using our modular architecture for news media analysis [15], [16]. Previously, this system has been successfully used for several media analysis studies in both news and social media, ranging from predicting flu levels from Twitter content [17], analyzing public mood from social media [18], large-scale analysis of topic, style and gender bias in news content [19], and detecting patterns in the news coverage of US elections [20].

News articles are labeled by our system as science articles (i.e. their subject matter is about science topics) in one of two ways. The first method is that all news articles

coming from an online news feed that was explicitly hand-annotated as "Science" or "Technology" inherit a science label, denoting them as science articles. Alternatively, we also automatically classify news articles into 15 different generic news categories, such as "Crime" or "Science", using Support Vector Machines (SVM) [21] trained for high precision on the New York Times [22] and Reuters corpora [23]. Any news articles receiving a positive label from the SVM trained for "Science" news is also included in this study as a science article. This includes science news from main stream newspapers and tabloids, along with science articles from more science focused news sources. We additionally restricted the science articles used for this study to those that are written in the English language.

We decided to focus our analysis only on science news articles, rather than general news, in an effort to ensure we are monitoring how the reporting of science has changed, rather than the general reporting of major events in the news. In total, our study covers $5,195,010$ science articles written in the English language, where the average length of a science article is 343 words.

III. Methodology

In order to examine how different science-based issues and events are framed by the mass media, we focused on analyzing the context of how different scientific concepts and associated actors (collectively referred to as 'items', including scientific topics, universities and diseases) are mentioned in the mass media. For each reference to one of these items, we compute a number of attributes of the item as found by analyzing the surrounding text.

Firstly, time series of the salience of each item were computed, along with the sentiment surrounding the item, demonstrating how the amount of attention and the opinions about the items changed during the period covered by this study.

Secondly, we mine the association rules between different items to discover how they are associated with each other based upon how often they co-occur in the same science articles. Thereby we can find which concepts are most closely associated to one another along with their most relevant actors, or how different actors interact with each other.

Thirdly, we extract Subject-Verb-Object (SVO) triplets for each of the items, allowing us to quickly discover what types of things are performing actions on the items, and also what actions are being performed by the items. This additionally allows for the analysis of the action clouds relative to an item, showing the collective actions taken or being taken on an item.

Data processing for the following methods took place using the Apache Hadoop framework[1], while data was stored and retrieved from MongoDB[2]. Time series data is separately generated using ElasticSearch[3].

A. Extracting References

References to the scientific topics, universities and diseases are extracted from the corpora of science articles by first compiling a list of the items which we wish to detect.

Scientific topics were generated using candidate lists of academic disciplines, along with scientific topics from Wikipedia [24], ranging from the natural sciences such as Biology and Chemistry to formal sciences such as Mathematics and Computer Science, and many others (Social, Applied, etc.). In total we tracked references for 677 different topic items.

The list of universities was compiled by taking the top 500 universities appearing in the "QS World University Rankings 2013" [25].

The diseases were collected using the lists of diseases and disorders available on Wikipedia [26], covering a wide range of diseases from cancers, infectious and non-communicable diseases as well as different disorders related to genetics, mental health etc. The disease list covers 4562 diseases overall.

Once the lists were compiled, we extracted the references to the items by running each science article through the ANNIE application [27] of the GATE architecture [28], performing tokenization, sentence splitting, part of speech tagging, and finally identifying references from the lists using the gazetteer.

B. Generating Time Series

We generate two types of time series, revealing first the amount of attention a given item receives and also the sentiment surrounding a given item over time. In this study, we resolve the time series to weekly segments, giving us 297 time points in our time series.

The attention focused on a particular item is calculated as the percentage of science articles in the corpus which contain a reference to the item. To generate a time series, we split the corpus into segments based upon the date of publication for each science article, with each segment containing all science articles in that time period. Using these segments, we can then calculate the relative frequency of science articles containing the item as the number of science articles containing the item over the total number of science articles in the segment, referred to as the relative frequency of an item at a given time.

Time series displaying the sentiment for a given item over the period covered by the science corpora can also be generated by first finding every occurrence of a reference to the item in the corpora. For each occurrence, the number of positive and negative sentences that contain it are calculated

[1]Apache Hadoop: http://hadoop.apache.org/

[2]MongoDB: http://www.mongodb.org/

[3]ElasticSearch: http://www.elasticsearch.org/

using Bing Liu's Opinion Lexicon [29] which contains 6800 polarized sentiment terms, and has been shown [30] to have a low level of disagreement with the MPQA lexicon [31], Harvard General Inquirer [32] and LIWC [33] sentiment lexicons.

A sentence containing a reference to an item is then defined as positive if it contains more positive terms than negative terms, and vice versa. The publication date for each sentence allows the aggregate sentiment score to be computed for every week covered by the corpora by taking the difference between the positive and negative sentences, normalized by the total number of sentences in the science articles that contain a reference to the item.

C. Mining Associations

Associations between the items were obtained by performing association rule mining using the FP-Growth algorithm [34]. Each science article is treated as a transaction, with each reference to an item being an object in the transaction. We mined up to a maximum of 1000 association rules for each item. The confidence in an association between two items gives us an estimate of the probability of seeing an associated item, given the target item has been mentioned.

From the association rules, we build up networks with items as the nodes, and the edges between nodes encoding the confidence in the association. This allows us to quickly visualize how probable it is that items will co-occur with each other in the same science articles.

D. Extracting Triplets and Action Clouds

Triplets are extracted from the science articles by first resolving co-references and performing anaphora resolution on the text, before running the Malt dependency parser [35] and generating a full parse of the text. From each parse, we then extract triplets that match the form Subject-Verb-Object, tracking how often they occur and in which science articles they appear.

We generate triplet networks for each item by collecting together all triplets where either the subject or the object match the item, displaying this as a network with each item represented by a node, with edges showing the action relating the adjacent items. The triplet networks are then pruned to remove noise, keeping only nodes which occur more than once as either a subject or object in the extracted triplets.

Action clouds are generated by aggregating together all the verbs from the triplets where a particular item forms the subject or object of the triplet. This allows two action clouds to be generated for each item, one for the actions performed by the item, and one for the actions performed on the item.

IV. Results

We give results for each of the methods detailed in the previous section, demonstrating their feasibility, focusing our attention on the topic of "Nuclear Power", showing how a big data approach to corpus analysis can reveal information about events and issues in the corpus using these methods.

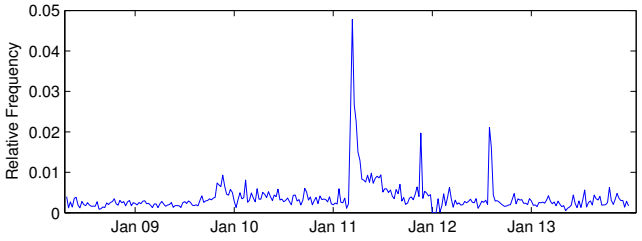

Figure 1. Relative frequency of the number of science articles mentioning 'Nuclear Power' between 1st May 2008 and 31st December 2013.

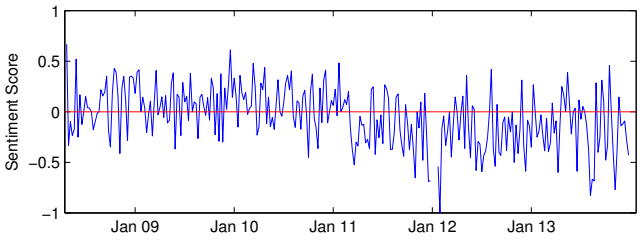

Figure 2. Normalized difference in the number of positive to negative sentences mentioning 'Nuclear Power' between 1st May 2008 and 31st December 2013.

A. Evolution of Attention and Sentiment

Monitoring the longitudinal evolution of the attention and sentiment surrounding issues and actors allows us to see which real-world events garner the attention of the media and provoke a sentimental reaction in its coverage, as well as those which did not. Perhaps not surprisingly, Figure 1 shows the large effect in terms of increased salience and therefore media attention that the Fukushima Daiichi incident brought on nuclear power for a period of a few months. Overall though, it does not suggest a prolonged effect on the attention given to the topic, with the average relative frequency of mentions returning to the same levels as before the incident.

Regarding the sentiment surrounding nuclear power in the past years, it had lost part of its negative stigma in the context of fighting climate change and improved safety of installations. Figure 2 supports this claim showing a relatively positive media coverage from 2008 to March 2011. However, the impact of the incident is evident in the shift from a positive to negative coverage, with a long lasting effect that does not appear to recover before the end of the period covered by this study.

B. Associations

Performing association rule mining on our items reveals information about how often items co-occur with one another during the media discourse, and in this case allows us to gain an idea of the most associated topics, universities

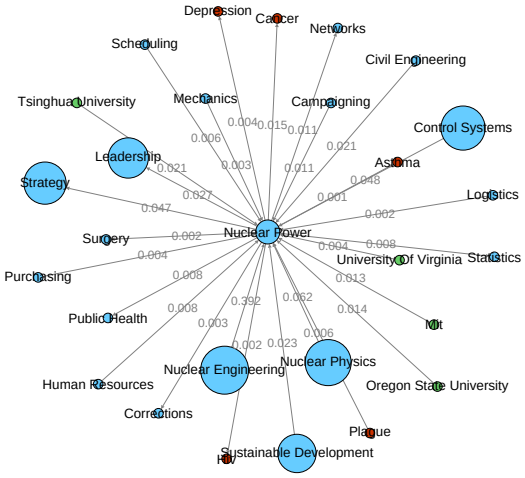

Figure 3. Associated universities (green), topics (blue) and diseases (red) found through association rule mining for 'Nuclear Power' before the Fukushima disaster. Edge weight denotes the confidence in the association.

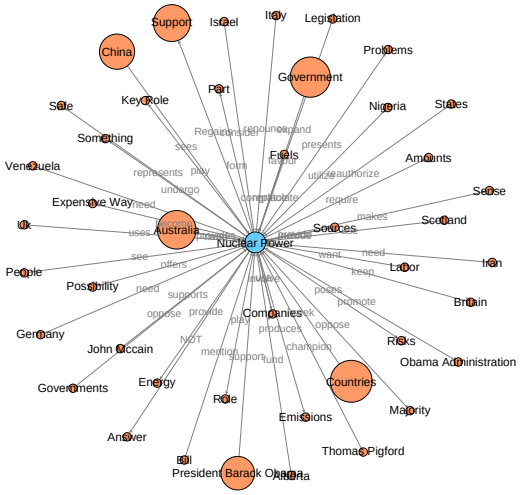

Figure 5. SVO triplet network showing the actors and actions affecting 'Nuclear Power' before the Fukushima disaster. Nodes represent subjects and objects in the SVO triplets, while edges show the verb relation between the subject and object of the triplet.

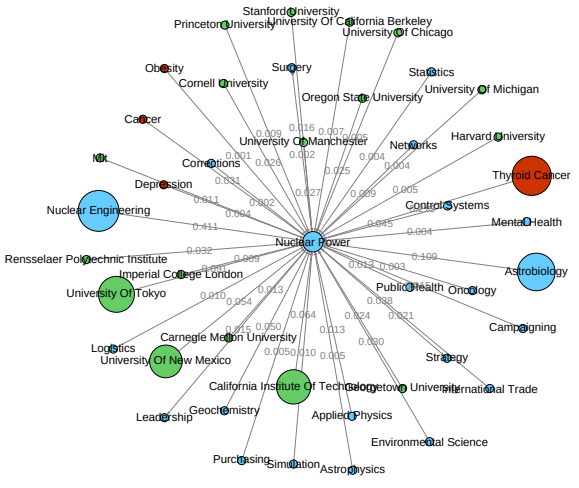

Figure 4. Associated universities (green), topics (blue) and diseases (red) found through association rule mining for 'Nuclear Power' after the Fukushima disaster. Edge weight denotes the confidence in the association.

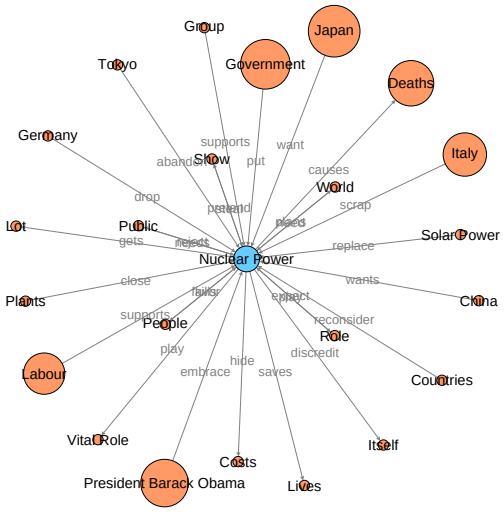

Figure 6. SVO triplet network showing the actors and actions affecting 'Nuclear Power' after the Fukushima disaster. Nodes represent subjects and objects in the SVO triplets, while edges show the verb relation between the subject and object of the triplet

Figure 7. Verbs in the SVO triplets where 'Nuclear Power' is the object before the incident, showing the actions happening to 'Nuclear Power'.

Figure 8. Verbs in the SVO triplets where 'Nuclear Power' is the object after the incident, showing the actions happening to 'Nuclear Power'.

and diseases to a particular topic.

We generated association networks showing the co-occurrences in the science articles for nuclear power before and after the incident, exposing the change in topics mentioned in relation to nuclear power. The edges of the networks express high co-occurrence between items in the same science articles and are weighted by the confidence of the association as described in Section III-C.

Before the Fukushima disaster, nuclear power was associated with other scientific topics and institutions and only marginally related to health risks such as cancer, plague, asthma, etc. as seen in Figure 3.

After the Fukushima incident, there is a substantial increase in associating nuclear power with cancer and in particular with thyroid cancer (a common radiation-induced form of cancer) as shown in Figure 4. This is due to the aftermath of the incident in Japan and a more frequent media discourse on the health hazards of nuclear power in which cancer is predominant. These findings are in line with what we described both in terms of the evolution of articles sentiment and the following networks of actors and actions.

C. Actions and Actors

Extracting SVO triplets and generating a triplet network for individual items allows us to map the range of issues related to each item, clearly displaying the key actors and entities affecting and being affected by the items in turn. By increasing the level of granularity and looking directly at the actions relating to nuclear power before and after the incident we can analyze how the media frames the topic in a different light.

Figure 5 shows the network of actors and actions linked to nuclear power before the incident and contains a number of policy actors and countries revealing the debate about nuclear power as a viable alternative to fossil based energy sources. Common frequent actors are countries or political figures because most articles reflect the debate taking place within countries about their energy supply needs, with the actions linking them also reflecting such a kind of discussion. This can be further seen in Figure 7, where the collective actions being applied to nuclear power are mostly centered around 'support', 'embrace' and 'need'.

However, after the Fukushima disaster the network of actors and actions changed (Figure 6). The biggest change is the introduction of the public as a very important actor and their views and feelings about nuclear power. The role and risks associated with nuclear power re-emerged as an element of the debate. Actions such as 'replace', 'reject' and 'abandon' become more prominent, as seen in Figure 8. Additionally, in contrast to pre-incident, actors' public reaction and acceptance of nuclear power gained a prominent role in its online news representation and we can guess that this is the reason of negative long-standing shift in the articles sentiments after the incident.

V. Discussion

Traumatic and dramatic events can have a profound impact on the way media represents an issue and in an age of global media, such effects can potentially have a wide reach. Traditionally, it has been very difficult to ascertain which events had an impact or not in the public sphere, and discerning if the impact is a long-lasting one. The level of attention, sentiment and framing (both in terms of SVO triplets and associated topics) in the media coverage of a technology can greatly affect its trajectory. The most common recent example is the case of biotechnology in Europe and its public opinion backlash that generated an embargo for GMO crops [36].

Actors and other issues associated to nuclear power have experienced a dramatic change from a largely positive one to a clearly negative climate. Such dynamics remind of the case of the Three Mile Island accident in 1979 that had considerable effect on the development of the nuclear industry[4].

Further analysis of important events and issues in the public sphere could also be augmented by the inclusion of sentiment analysis of data from social media sources, allowing for a more direct method of gauging the public

[4]http://en.wikipedia.org/wiki/Three_Mile_Island_accident#Effect_on_nuclear_power_industry

reaction to particular issues. Similarly, associations and the key actors and actions could also be extracted from social media data for a fuller picture of how events and issues unfold in the public eye.

VI. Conclusion

Our findings reveal an insight into the change of framing and sentiment associated with the global media reporting of nuclear power following the nuclear disaster in Fukushima, Japan in March 2011. Before the incident, nuclear power had a relatively positive sentiment in the media, typically framed in terms of playing a key role in the ongoing debate within countries about managing their energy supply needs. Following the incident however, there is a negative shift in the sentiment surrounding nuclear power, with the debate drifting towards the perceived risks of nuclear power and links with thyroid cancer.

While we cover a global selection of science articles on the topic of nuclear power, a more fine grained approach, focusing specifically on the debate playing out within individual countries would be of interest, and would allow for findings to be compared against opinion polls carried out by surveys.

The methodology implemented in this paper presents a comprehensive way to monitor critical events and their media ripple effects that can be potentially applied to any publicly relevant issue. Big data provides a unique opportunity to map, monitor and study public sphere dynamics with a global and longitudinal approach revealing the true 'long tail' of events. In the past, previous media monitoring methodology based on human coding did not fully allow to detect and distinguish such effects. The innovative character of these techniques opens up new possibilities in social scientific research.

Acknowledgment

Nello Cristianini, Thomas Lansdall-Welfare and Saatviga Sudhahar are supported by the EU projects Complacs and ThinkBIG. Giuseppe A. Veltri is supported by the Development Fund of the Social Science College at the University of Leicester.

References

[1] S. Moscovici and G. Duveen, *Social representations: Explorations in social psychology*. Polity Press Cambridge, 2000, vol. 41.

[2] G. A. Veltri, "Viva la nano-revolución! a semantic analysis of the spanish national press," *Science Communication*, vol. 35, no. 2, pp. 143–167, 2013.

[3] M. E. McCombs and D. L. Shaw, "The evolution of agenda-setting research: twenty-five years in the marketplace of ideas," *Journal of communication*, vol. 43, no. 2, pp. 58–67, 1993.

[4] M. McCombs and S. I. Ghanem, "The convergence of agenda setting and framing," *Framing public life: Perspectives on media and our understanding of the social world*, pp. 67–81, 2001.

[5] G. Gerbner and L. Gross, "Living with television: The violence profile," *Journal of communication*, vol. 26, no. 2, pp. 172–194, 1976.

[6] G. E. Lang and K. Lang, *The battle for public opinion: The president, the press, and the polls during Watergate*. Columbia University Press New York, 1983.

[7] S. S. Fahmy, W. Wanta, T. J. Johnson, and J. Zhang, "The path to war: Exploring a second-level agenda-building analysis examining the relationship among the media, the public and the president," *International Communication Gazette*, vol. 73, no. 4, pp. 322–342, 2011.

[8] P. Slovic, M. Layman, and J. H. Flynn, *What Comes to Mind when You Hear the Words" nuclear Waste Repository"?: A Study of 10,000 Images*. State of Nevada, Agency for Nuclear Projects/Nuclear Waste Project Office, 1990.

[9] A. Boholm, "Comparative studies of risk perception: a review of twenty years of research," *Journal of risk research*, vol. 1, no. 2, pp. 135–163, 1998.

[10] W. A. Gamson and A. Modigliani, "Media discourse and public opinion on nuclear power: A constructionist approach," *American journal of sociology*, pp. 1–37, 1989.

[11] S. R. Weart, *Nuclear fear: A history of images*. Harvard University Press, 1988.

[12] G. A. Veltri and A. Suerdem, "Worldviews and discursive construction of gmo-related risk perceptions in turkey," *Public Understanding of Science*, vol. 22, no. 2, pp. 137–154, 2013.

[13] S. Ogden, "The language of agricultural biotechnology terminate or be terminated," *Organization & environment*, vol. 14, no. 3, pp. 336–340, 2001.

[14] S. K. Yeo, M. A. Cacciatore, D. Brossard, D. A. Scheufele, K. Runge, L. Y. Su, J. Kim, M. Xenos, and E. A. Corley, "Partisan amplification of risk: American perceptions of nuclear energy risk in the wake of the fukushima daiichi disaster," *Energy Policy*, vol. 67, pp. 727–736, 2014.

[15] I. Flaounas, T. Lansdall-Welfare, P. Antonakaki, and N. Cristianini, "The anatomy of a modular system for media content analysis," *CoRR*, vol. abs/1402.6208, 2014.

[16] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie, and N. Cristianini, "NOAM: News Outlets Analysis and Monitoring System," in *SIGMOD 2011*. ACM, 2011, pp. 1275–1278.

[17] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter," *Machine Learning and Knowledge Discovery in Databases*, pp. 599–602, 2010.

[18] T. Lansdall-Welfare, V. Lampos, and N. Cristianini, "Effects of the recession on public mood in the uk," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. ACM, 2012, pp. 1221–1226.

[19] I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "Research methods in the age of digital journalism: Massive-scale automated analysis of news-contenttopics, style and gender," *Digital Journalism*, vol. 1, no. 1, pp. 102–116, 2013.

[20] S. Sudhahar, T. Lansdall-Welfare, I. Flaounas, and N. Cristianini, "Electionwatch: detecting patterns in news coverage of us elections," in *Proceedings of EACL*. Association for Computational Linguistics, 2012, pp. 82–86.

[21] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[22] E. Sandhaus, "The new york times annotated corpus," *Linguistic Data Consortium, Philadelphia*, vol. 6, no. 12, p. e26752, 2008.

[23] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *The Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[24] Wikipedia, "List of academic disciplines and sub-disciplines," 2014. [Online]. Available: http://en.wikipedia.org/wiki/List_of_academic_disciplines_and_sub-disciplines

[25] Q. Q. S. Limited, "Qs world university ranking 2013 — top universities." [Online]. Available: http://www.topuniversities.com/university-rankings/world-university-rankings/2013

[26] Wikipedia, "Lists of diseases," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Lists_of_diseases

[27] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[28] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6)*, 2011. [Online]. Available: http://tinyurl.com/gatebook

[29] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[30] C. Potts, "Sentiment symposium tutorial: Lexicons - relationships," 2011. [Online]. Available: http://sentiment.christopherpotts.net/lexicons.html\#tab:lexicon_disagreement

[31] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[32] P. Stone, D. C. Dunphy, M. S. Smith, and D. Ogilvie, "The general inquirer: A computer approach to content analysis," *Journal of Regional Science*, vol. 8, no. 1, pp. 113–116, 1968.

[33] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, p. 2001, 2001.

[34] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *ACM SIGMOD Record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.

[35] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi, "Maltparser: A language-independent system for data-driven dependency parsing," *Natural Language Engineering*, vol. 13, no. 02, pp. 95–135, 2007.

[36] G. Gaskell and M. W. Bauer, *Biotechnology 1996-2000: the years of controversy*. Science Museum London, 2001.