University of Colorado Denver
School of Public Affairs

November 2014

Automated Analysis of News Articles on Hydraulic Fracturing in Colorado, New York, and Pennsylvania

Produced by the School of Public Affairs at the University of Colorado Denver

Authors
Benjamin Blair, Research Fellow
Chris Weible, Associate Professor
Tanya Heikkila, Associate Professor
Darrick Evensen, Visiting Assistant Professor, Oberlin College
Sam Gallaher, Doctoral Candidate

Acknowledgements

This research was funded by the Alfred P. Sloan Foundation, though the research design and results are the authors' alone.

Citing this Summary Report

Blair, Benjamin, Weible, Christopher M., Heikkila, Tanya, Evensen, Derrick, and Gallaher, Samuel. 2014. "Automated News Article Analysis on Hydraulic Fracturing in Colorado, New York, and Pennsylvania." Published November 2014 by the School of Public Affairs University of Colorado Denver.

Questions, Comments, and Requests for More Information

For questions, comments, concerns, and feedback regarding this research project please contact the following:

Tanya Heikkila
Associate Professor
School of Public Affairs
University of Colorado Denver
1380 Lawrence Street, Suite 500
Denver, CO 80217
Phone: 303-315-2269
Fax: 303-315-2229
Email: Tanya.Heikkila@ucdenver.edu

Chris Weible
Associate Professor
School of Public Affairs
University of Colorado Denver
1380 Lawrence Street, Suite 500
Denver, CO 80217
Phone: 303-315-2010
Fax: 303-315-2229
Email: Chris.Weible@ucdenver.edu

Table of Contents

Executive Summary

This report presents the findings of different approaches to media analysis of newspaper articles that focus on unconventional oil and gas development inclusive of hydraulic fracturing in New York, Pennsylvania, and Colorado. A total of 2,114 articles are evaluated (1,037 from NY and PA from 2007 to 2011 and 1,077 from CO from 2008 to 2013) from seven different newspapers: Binghamton Press and Sun Bulletin and Elmira Star-Gazette (NY), Boulder Daily Camera, Colorado Springs Gazette, and Denver Post (CO), and Scranton Times-Tribune and Williamsport Sun-Gazette (PA). The objectives of this report are twofold.

The first objective is to *compare the results of manual coding of media articles on hydraulic fracturing issues from newspapers in New York and Pennsylvania to the automated coding results using the latent Dirichlet allocation (LDA) on the same articles*. The purpose is to see if the automated coding can be used as a valid alternative to human coding, particularly when researchers are interested in coding a large number of articles. The findings show that, when comparing the automated coding results to the manual coding results, the inter-rater reliability (Cohen's kappa = 0.21) was considered "fair" and the percent agreement for economic, environmental, and social topics is approximately 61%. In addition, some of the articles that were manually coded did not contain any of the codes which the researchers were searching and when these articles were removed from the analysis, the percent agreement increased to 64% (Cohen's kappa = 0.25, classified as "fair"). The results also showed the accuracy for environment and economic topics to be 65% and 67%, respectively, which is similar to automated coding results from comparable studies (Grimmer and Stewart, 2003).

The second objective is *to evaluate three Colorado newspapers and assess the differences in the presentation of topics from the coverage of hydraulic fracturing based on the political orientation of the newspapers*. The findings showed that the more conservative leaning newspapers used a greater proportion of economic words, while the liberal newspapers used a greater proportion of public health and environmental damage words. Additionally, we find that the significant differences can be detected in the dimension, the co-occurrence of topics, in the newspapers evaluated.

This report provides a rare validity check on automated approaches of news media based on manual coding of the same article and offers modest support for using a LDA approach to understand the propensity of different topics across a large number of articles at a broad level. Further research is needed to advance automated coding methods to increase the accuracy and precision and establish the best practices for the coding the valence of topics. The LDA approach can also be used to distinguish differences in patterns and phrasing between liberal and conservative newspapers. We recommend further research to more accurately predict the topics and potential informational bias in the news media.

1.0. Introduction

Freedom of the news media is one of the hallmarks of any democracy and the media has the ability to communicate various forms of information across mass publics. For example, the media have a crucial responsibility as a source of scientific information for citizens as well as framing or stories that depict different interpretations of events (Carvalho, 2007; Shanahan et al., 2011). Additionally, the general public acquires most of its knowledge about science from the news media (Wilson, 1995; Boykoff and Boykoff, 2004). In the context of oil and gas development involving hydraulic fracturing, for example, a recent survey of randomly selected households in New York and Pennsylvania found that mass media were the most used source for information (Stedman et al., 2012). Furthermore, the information and ideas held about hydraulic fracturing can strongly influence the discussion of this issue and the types of regulation viewed as appropriate (Evensen et al., 2014). Comparing and developing methodological approaches for analyzing news media and to analyze issues, such as potential bias in coverage, is needed. Given the central role of the news media in relating information to mass publics, this report explores two different approaches for analyzing such coverage:

1. *To compare the results of human and automated coding of media text.* Capturing systematic differences in the news media has been accomplished in numerous studies focusing on complex scientific issues, such as global warming (Boykoff and Boykoff, 2004). However, traditional manual or human coding methods are time consuming and financially costly; therefore, attempting to automate the coding of documents has been a common endeavor among many media scholars (Van Holt et al., 2012; Mcnamara, 2005). In this report, we attempt to mirror the human coding results of media reporting through the use of automated text analysis, along with developing methods to further evaluate the news media using automated techniques. Additionally, the news media is often accused of biased or selective coverage of events based on ideology. For example, recent public opinion surveys show public trust in the news media has declined over time (e.g. Gallup Poll, 2013). This first objective seeks to see if automated coding can reliably and validly substitute for human coding, which is costly and time consuming.

2. *To conduct automated media analysis of liberal and conservative newspapers to assess differences in coverage on the hydraulic fracturing issue.* This second objective uses automated coding in two ways to compare newspaper coverage. This analysis also includes a comparison of editorials and regular news articles for liberal and conservative newspapers.

The recent increase in the deployment of hydraulic fracturing for oil and natural gas development has led to one of the most contentious issues in environmental and energy policy. The goal of this report is to explore different approaches for understanding media coverage of the debate surrounding hydraulic fracturing ("fracking") and other oil and natural gas development in Colorado, New York, and Pennsylvania from 2007 to 2013. Articles from seven newspapers were evaluated: Binghamton Press and Sun Bulletin and Elmira Star-Gazette from

New York (NY), Boulder Daily Camera, Colorado Springs Gazette, and the Denver Post from Colorado (CO), and Scranton Times-Tribune and Williamsport Sun-Gazette from Pennsylvania (PA). In analyzing articles in these seven newspapers, we focus on the issues or "topics" that have emerged across debates associated with hydraulic fracturing, which include: environmental concerns, economic harm or benefits, potential harm to public health, and social issues.

1.1. Brief Hydraulic Fracturing Overview and Presentation of Topics

Hydraulic fracturing, also referred to as "fracking," "fracing," or "hydrofracking," is the process of pumping a mixture of water, sand or similar material, and chemical additives, under high pressure, into vertically or horizontally drilled wells that allows for natural gas and oil recovery. The development of unconventional sources of oil and natural gas using hydraulic fracturing has significant controversy. Opponents of hydraulic fracturing argue that it may harm public and environment health while supporters argue that it is a critical step to ensure domestic energy security and that it offers a transition away from coal based energy production (Engelder, 2011; Howarth, Ingraffea, 2011).

This study focuses on the media framing of hydraulic fracturing in New York, Pennsylvania, and Colorado. These three states are governing hydraulic fracturing differently. Hydraulic fracturing expanded rapidly in Pennsylvania staring in 2007 and was the initial epicenter of the conflict in the United States. New York, in contrast, has had a de facto moratorium on the development of hydraulic fracturing since 2008. Colorado has seen a rapid development of hydraulic fracturing but not as much as Pennsylvania. For summaries of hydraulic fracturing in these three states, readers are directed to the following sources (Pierce et al., 2013 for Colorado; Heikkila et al., 2014 for New York; Evensen et al., 2014 and Brown et al., 2013 for Pennsylvania). Based on the results from these articles and reports, the media analyses in this report are established from a set of primary topics commonly found in the debates on this issue in the three states: economic, environmental, social, and public health.

Concerns regarding public and environmental health have been raised from hydraulic fracturing due to the potential for surface water contamination (Warner et al., 2013), ground water contamination (Osborn et al., 2011), and air contamination (McKenzie et al., 2012), but more research is needed to clarify the magnitude of the potential risks (Shonkoff et al., 2014; Adgate et al., 2014). Likewise, the economic benefits and impacts from hydraulic fracturing are widely debated; where some say the economic benefits outweigh the risk, while others say the cost to the environment outweighs the benefits from the natural gas (Engelder, 2011; Howarth, Ingraffea, 2011; Kinnaman, 2011). Furthermore, social concerns, including noise, traffic, dust, crime rate, etc. are concerns commonly presented in the hydraulic fracturing discussion (Evensen et al., 2014; Jacquet, 2014). The combination of these technical disputes and goal conflicts cause hydraulic fracturing to be one of the more contentious issues in environmental and energy policy.

2.0. Methods

We obtained the Binghamton Press and Sun Bulletin (NY) articles through ProQuest, the Boulder Daily Camera (CO) and the Denver Post (CO) articles through LexisNexis, the Colorado Springs Gazette (CO) articles through Newspaper Archive, the Elmira Star-Gazette (NY) articles through a university library subscription to that newspaper, the Scranton Times-Tribune (PA) articles through Access World News, and the Williamsport Sun-Gazette (PA) articles from the free online archives of that newspaper.

For the New York and Pennsylvania articles, the search phrase was "Marcellus Shale," which allowed the majority of the articles focusing on issues of share gas extraction and development to be identified. The search dates spanned from January 1, 2007 To December 31, 2011. For the Colorado articles, we used key search terms of "hydraulic fracturing," "fracking," "fracing," and "shale gas." The search dates spanned from January 1, 2008 To December 31, 2013. The Colorado articles were briefly reviewed to ensure the articles included a substantial section regarding hydraulic fracturing. A total of 2,114 articles were evaluated. Of these, 246 were from Binghamton (NY), 462 from Boulder (CO), 509 from Denver (CO), 127 from Colorado Springs (CO), 229 from Elmira (NY), 268 from Scranton (PA), and 291 from Williamsport (PA).

2.1. LDA Summary

The substantial time commitments and corresponding costs of analyzing large collections of texts have prevented many researchers from analyzing such collections (Grimmer and Stewart 2013). To overcome this hurdle, automated text analysis methods have been developed. This report utilizes latent Dirichlet allocation (LDA) as an automated text analysis method. LDA is a data clustering technique where groups of words are clustered together into topics and the proportion of the text is then assigned to these topics. Of course, these clusters are not readily apparent in a single news article or hundreds to thousands of news articles. Therefore, the principal contribution of the LDA is to help identify one or more of these latent clusters of topics where a topic is determined probabilistically by the co-occurrence of words (Blei et al., 2003).

The LDA can estimate words within topics and the co-occurrence of topics within documents simultaneously, with the goal of inferring the structure of topic proportions within documents. More simply, the LDA assigns each individual document a topic proportion to different latent topic clusters based on the co-occurrence of words within a document. Hence, the output of a LDA proportions can be used to assess the extent that one or more documents emphasize certain topic clusters more than others. In addition, supervised LDA methods can use predetermined dictionaries and topic clusters to focus the LDA methods to test more specific hypotheses. In this way, the approach offers an opportunity to define hundreds to

thousands of documents into topics that offer insight into the general framing of the text under study.[1]

2.2. LDA Software

LDA is an automated approach that can be used within multiple software packages. In this research, Mallet version 2.0.7 (McCallum, 2002) was used for the LDA. To support the LDA, Automap version 3.0.10 (Carley et al., 2012) was used for stemming (using "environ" for "environment" and "environmental")) and to create list of top words and their corresponding frequency of use. The thesaurus function was also used in Automap as outlined in section 2.4. Additionally, IBM SPSS Statistics 22.0 was used for to complete the ANOVA calculations, regressions; and the Cohen's kappa values used for inter-relater reliability.

The LDA was used in two different ways, the dictionary method and the Automap method, which are summarized in the next two sections.

2.3. Dictionary Method LDA

An automated dictionary method LDA was used to compare the automated results with the human coding found in Evensen et al. (2014). The dictionary method requires scholars to identify words and place them into separate issues or topics (Grimmer and Stewart 2013). To evaluate the topics, words were placed into one of the three coded topics (environmental, economic, and social), when applicable. For example, the words "air" and "water" were both placed in the "environment" topic, which are words that are in agreement with the coding scheme used by Evensen et al. (2014). Words were selected from the list of words in Appendix 1 of the Evensen et al. (2014) paper. Next additional words were then identified inductively and these included high frequency words (appearing in the top 2000 words from all articles, based on count).

Slightly different dictionaries were developed for the New York/Pennsylvania and Colorado analyses due to the different themes coded in previous work evaluating these States (Evensen et al., 2014; Heikkila et al., 2014). Three primary topics were evaluated via content analysis for the New York and Pennsylvania newspapers: economic, environmental, and social. Four primary topics were evaluated via content analysis for the Colorado papers: economic, environmental, environmental damage, and public health. The methods for the human coding of topics are explained in great detail in the Evensen et al. (2014) paper.

Stemming refers to the process of removing the suffix of words to equate similar words. For example, stemming removes 'ed' or 'ing' from words to combine similar words within the dictionary. The Porter stemming algorithm (Porter, 1980) was employed in Automap because

[1] For further information on LDA process and automated text analysis, we recommend the papers by Blei et al. (2003) and Grimmer and Stewart (2013).

of its moderate approach to word simplification (Grimmer and Stewart, 2013). Stemming was used in the New York and Pennsylvania articles to reduce words to their base or stem.

The LDA was then used to assign each article a proportion for each of these predetermined topics. An example of the code for the dictionary LDA method is displayed in Appendix A.

2.3. Automap Method LDA for Colorado

To further evaluate the abilities of the LDA, we explore the combination of the functionality provided by the Automap software package (Carley et al., 2012) and the LDA. This was employed to evaluate the simultaneous occurrence of multiple topics within the documents. In general, the Automap method was employed to further advance an understanding of co-occurrence of topics presented.

In this work, a dimension is defined as a cluster of topics. For example, the topics "environment," "economic," and "public health" can be combined into a dimension and various dimensions can be evaluated at the document level. Therefore, the co-occurrence of multiple topics will create a dimension and multiple words create a topic. Similar to the dictionary method, our thesaurus was developed via both an a priori and inductive approach. Then, after replacing the words into topics, all unidentified words were replaced with a stop word that is excluded from the LDA results. A traditional LDA was then completed on the modified documents using Mallet. An example of the code is displayed in Appendix B

3.0. Results and Discussion

3.1. Dictionary Method LDA for New York and Pennsylvania

The dictionary method LDA was used to place words into predetermined topics based on the human coding results from Evensen et al. (2014). The words were stemmed and, when applicable, were placed into one of three primary topics, as shown in Table 1.

Table 1. Primary topics and stemmed words within each topic based on the Evensen et al. (2014) human coding.

Topic	Words
Economy	royal cheap businessman tourist jobcreat jobless tourism rental tour taxpay invest economi royalti monei budget cost revenu properti busi econom job tax leas industri compani
Environment	fisheri drinkingwat greenhous habitat cleanup wildlif soil green solid radioact environmentalist fish lake watersh air stream methan pollut spill clean treatment forest remedi contamin wastewat drink environ river environment
Social	noisi crime roadwai dust nois light street independ infrastructur traffic import park hous road

Three LDA runs were completed, where each topic was compared with just one of the other topics (Economy and Environment, Economy and Social, Environment and Social). This method was selected to overcome the challenge of articles that had a very high co-occurrence of a single topic, which increased the complexity of evaluating whether the other two topics were present in the article. For example, when an article had a proportion of environment of 0.7, the difference between the remaining proportions was too small to evaluate the presence or absence of a topic. Once the LDAs were completed on the 1,037 articles from New York and Pennsylvania, every other article was selected for the model fitting step. In this step, cutoff values were evaluated for accuracy and precision, as defined in equations 1 and 2 (Grimmer and Stewart, 2013).

$$\text{Accuracy} = \frac{\text{Number of correctly classified documents}}{\text{Total number of documents}} \quad (1)$$

$$\text{Precision} = \frac{\text{Number of documents correctly classified}}{\text{Total number of documents the machine classifies correctly}} \quad (2)$$

Accuracy is the proportion of correctly classified documents and precision is the probability that the machine made the right guess (Grimmer and Stewart, 2013). Using the accuracy and precision for the three topics, cutoff values to be used in the validation step were determined by selecting the values with a minimum deviation from 1.0 for the accuracy and precision values when the proportion was converted to whether a topic was present. The

values are shown in Table 2, and the model fitted values that were selected for each topic are bold.

Table 2. Accuracy and precision for the three topics used for model fitting. Values in bold were the values selected for the model fit.

	Environment		Economic		Social	
Cutoff Value	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
0.1	0.68	0.68	0.63	0.63	0.33	0.33
0.15	0.68	0.68	0.63	0.63	0.33	0.33
0.2	0.68	0.68	0.63	0.63	0.33	0.33
0.25	0.68	0.69	0.63	0.63	0.33	0.33
0.3	0.68	0.7	0.63	0.63	0.34	0.36
0.35	0.67	0.74	0.63	0.63	0.38	0.42
0.4	**0.66**	**0.82**	**0.64**	**0.66**	0.46	0.69
0.45	0.59	0.96	0.61	0.71	**0.57**	**1.73**
0.5	0.49	1.51	0.5	0.87	0.66	13.23
0.55	0.38	3.62	0.46	1.36	0.66	68.6
0.6	0.33	13.31	0.42	2.84	0.67	173

Cutoff proportion values are used to fit the model prior to the validation step. Cutoff proportion values were selected at 0.4 for environment and economic topics and a cutoff proportion value of 0.45 was selected for social topics.

To convert a proportion to the presence or absence of a topic, the topic was considered present when the topic proportion was greater than the cutoff value for both LDA runs. The topic was considered absent when the topic proportion was below the cutoff value for either LDA run. For example, if the proportion for the environment topic was 0.5 when evaluated with economic and 0.45 when evaluated with social, the theme is considered present. In contrast, if the proportion for the environmental topic was 0.5 when evaluated with economic and 0.35 when evaluated with social, the theme is considered absent.

Once the model fitted cutoff proportion values were selected, the values were used in the second stage of the model, the validation step. Here, the model fitted cutoff proportion values for each topic were then applied to the remaining 518 New York and Pennsylvania articles for validation. The results are shown in Table 3.

Table 3. Accuracy and precision for validation step (n=519 articles).

	Environment	Economic	Social
Accuracy	0.61	0.62	0.55
Precision	0.80	0.65	1.70

A slight decrease was observed in accuracy when using the validation set compared to the model fitting articles. However, the accuracy values are comparable to accuracy levels for complex documents found in other studies (Grimmer and Stewart, 2013; Hopkins and King 2010).

An issue with the coding scheme is when articles are included in the analysis, but do not include any of the topics based on the human coding results. These articles are often about hydraulic fracturing, or just mention the process, but do not include any coded topics. When the articles without any human coding are removed from the analysis, the accuracy increases for both the model fitting and validation articles. These results are summarized in Table 4.

Table 4. Accuracy and precision for validation articles (n=519 articles) after the documents with no human coded topics were removed.

	Model Fitting			Validation		
	Environment	Economic	Social	Environment	Economic	Social
Accuracy	0.70	0.69	0.57	0.65	0.67	0.54
Precision	0.86	0.71	1.74	0.84	0.71	1.72

The inter-rater reliability was assessed using Cohen's kappa, which evaluates the degree of agreement among raters. The Cohen's kappa was found to be 0.213 across the model fitting and validation set, which would be considered "fair" inter-rater reliability (Viera and Garrett, 2005). However, for human coding documents with multiple human coders, a Cohen's kappa of 0.9 or greater is often achieved. Although identifying articles without topics is not possible without human coding, it demonstrates an area of improvement for further research. In addition, the Cohen's kappa increased to 0.25 across all three topics after removing the articles without any topics human coded.

This approach uses a model fitting and validation step to convert topic proportion values to either the presence or absence of a theme. Although this requires extensive human coding of documents prior to the use of automated coding, this approach allows for the analysis of the accuracy and precision of the modeled results. The results show the accuracy for environment and economic topics to be as high as 0.65 and 0.67, respectively. The social theme had a lower accuracy, at 0.54, which is likely due to the overall small number of documents that include this theme from the human coding results or with the theme being mentioned in passing in articles that focused primarily on environmental or economic considerations.

3.2. Dictionary Method LDA for Colorado

Two newspapers were selected in Colorado for the dictionary method: the Boulder Daily Camera and the Colorado Springs Gazette. These newspapers were selected due to the varying political endorsements and their close proximity to active hydraulic fracturing sites. The Boulder Daily Camera is considered liberal having endorsed the Democratic candidate for the past three Presidential elections (Boulder Daily Camera, 2012; Washington Post, 2004). The Colorado Springs Gazette is considered conservative where they endorsed Mitt Romney in 2012 and did not endorse a candidate in 2008 and 2004 (New York Times, 2012). We might expect, for example, based on the ideology of the newspapers that their coverage of this issue might differ with the Boulder Daily Camera being more concerned about environmental and public health issues and the Colorado Springs Gazette more concerned about economic issues.

Using the primary topics that have been identified in our previous work (e.g. Heikkila et al., 2014), we created four topics to evaluate the differences between the Colorado Springs Gazette and the Boulder Daily Camera. These topics included public health, economics, environment, and environmental damage, as summarized in Table 4. The results from this analysis are shown in Table 6.

Table 5. Dictionary Method LDA topics and words within each topic.

Topic and Topic Number				
	Public Health **0**	**Economic** **1**	**Environmental Damage** **2**	**Environment** **3**
Words within topic	health	economic	waste	water
	concerns	worker	emissions	environment
	safety	work	concerns	air
	drinking	unemployment	damage	environmental
	public	tourism	toxic	groundwater
	water	taxes	radioactive	wildlife
	endocrine	subsidies	carbon	wilderness
	supplies	spending	risks	watersheds
	benzene	sales	risk	watershed
	traffic	sale	warming	sustainable
	radon	royalty	uranium	surface
	exposure	royalties	spills	streams
	children	revenues	spill	soil
	cancer	revenue	sound	sea
	arsenic	recession	sewage	rock
	disruption	property	seismic	rivers
	safe	profits	radiation	river
	residents	production	quakes	quality
	clean	prices		pond

Table 6. Topics and corresponding proportions for Boulder Daily Gazette and Colorado Springs.

Topic	Boulder	Colorado Springs
Economic**	0.266	0.310
Environment*	0.249	0.264
Environmental Damage**	0.229	0.206
Public Health**	0.257	0.220

*p<.05 from independent samples t-test, equal variance not assumed
**p<.001 from independent samples t-test

The results show that the economic topic had a higher proportion per article, on average, in the coverage presented by Colorado Springs than in Boulder. In contrast, articles in Boulder were more likely to present environmental damage and public health topics. The difference between the presentation of environmental topics was statistically significant, but to a lesser extent than the other topics. Overall, the framing of the issue differed and would generally agree with the expectations based on their political endorsements.

3.3. Automap Method LDA for Colorado

This section compares news articles and editorial articles from three Colorado newspapers using the Automap method LDA. The Automap method LDA replaces words into predetermined topics and then the co-occurrence of these topics is found to create the dimensions. The three Colorado newspapers analyzed were the Boulder Daily Camera, the Colorado Springs Gazette, and the Denver Post. These sources were selected due to the varying political endorsements and their close proximity to active hydraulic fracturing sites. As summarized in the previous section, the Boulder Daily Camera is considered liberal and the Colorado Springs Gazette is considered conservative based on their presidential endorsement history. The Denver Post has a mix of presidential endorsements over recent elections, though leans democratic over time. The Denver Post endorsed Barack Obama in 2012 and 2008 and George W. Bush in 2004 (Associated Press, 2012). Additionally, being the state wide newspaper, we expect the Denver Post to be less liberal than Boulder and less conservative than Colorado Springs.

In this section, we evaluate the simultaneous occurrence of multiple topics within an article in Colorado. This was accomplished by evaluating the co-occurrence of topics within the text documents, but the LDA would determine the grouping of the topics, to create a dimension. Similar to the dictionary method, our thesaurus was developed via both an a priori and inductive approach. The topics and definitions used are as follows:

List of topics and definitions:

- **Economic:** Broad topics including words used to reference economic tools or the status of the economy
- **Energy:** Referencing energy, but not specific to type of energy used
- **Energy, Not Oil and Gas:** Specific to energy created using coal, nuclear, renewable, solar, or wind
- **Energy, Oil and Gas:** Specific to energy created using oil and gas
- **Energy Security:** Refers to energy dependence and imports/exports
- **Environment:** Any reference to the environment, ecosystem, or wildlife
- **Environmental Damage**: Words reference harm to the environment, such as contaminated, earthquakes, and toxic waste
- **Federal Policies:** The name of specific US Federal Polices
- **Fracking:** Using the words “frack”, “fracking”
- **Hydraulic Fracturing:** Using the words “Hydraulic Fracturing” or ‘Hydrofracking”
- **Moratorium:** Referencing the moratorium or banning of hydraulic fracturing
- **Process**: Specific to a process used in oil and gas mining
- **Public Health:** Includes references to public health issues including pollutants, toxicity endpoints, and references to drinking water contamination

- **Regulation:** Includes rules, policies, taxes, and leases
- **Research:** Referencing science, universities, institutes, professors, and reports
- **Rights:** Mineral and property rights
- **Setback:** Specifically mentioning setbacks

Preliminary research and interviews suggested the words used to describe the process of hydraulic fracturing were part of the framing of different policy actors. Because of this, we have defined two subtopics. One includes the words “frack” or “fracking” and the other includes “hydraulic fracturing” or “hydrofracking”. During the time period selected, advocates promoting oil and gas development would use “hydraulic fracturing” or “hydrofracking” compared to advocates opposed to oil and gas development who used the word “fracking” or “frack” when naming the process.

The results of the LDA, with a target number of 10 dimensions and 3 topics per dimension, are shown in Table 7. The number of dimensions and topics per dimension were established via sensitivity analysis: the numbers of dimensions evaluated ranged from 5 to 20 and the number of topics per dimension was evaluated between 2 and 10. Ten dimensions and 3 topics per dimension were selected due to the dimensions have a more coherent meaning than the results using the other configurations. Not all of the topics from the list above were selected by the LDA, due to the low comparative co-occurrence compared to the other topics.

Table 7. Dimension numbers and topics within each dimension from Automap LDA.

Dimension Number	Topics within each Dimension
0[a]	fracking publichealth hydraulicfracturing
1[b]	process environment economic
2[a]	moratorium fracking regulation
3[a]	environmentaldamage environment publichealth
4[c]	regulation rights economic
5[d]	research hydraulicfracturing process
6[e]	energyoilandgas economic process
7[d]	environment rights fracking
8[b]	economic regulation rights
9[a]	energynotoilandgas energyoilandgas energy

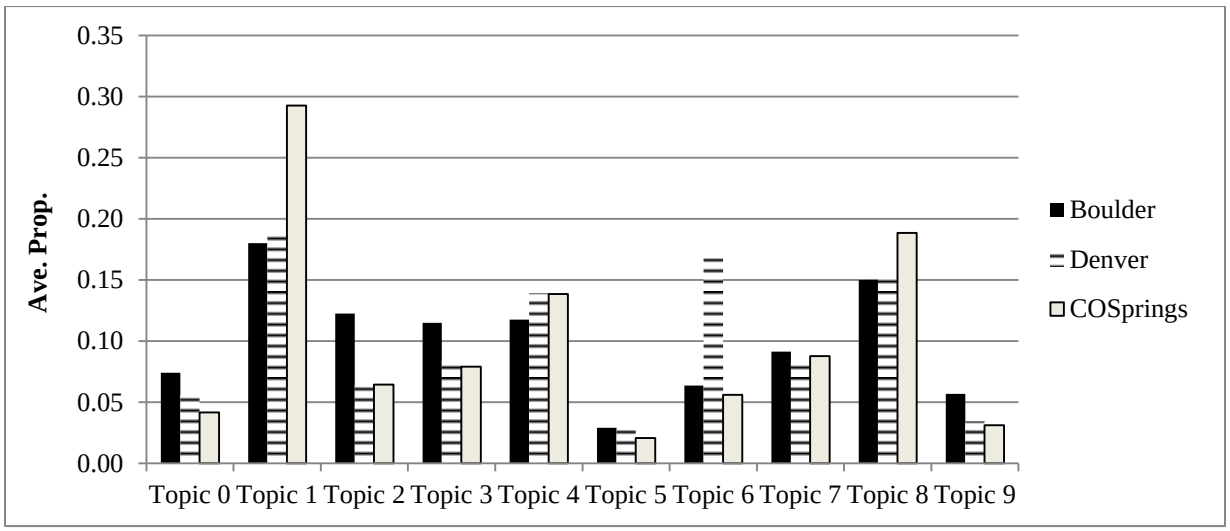

Figure 1. Average proportion of ten topics for the Boulder Daily Camera, Denver Post, and Colorado Springs Gazette. Topics are summarized in Table 7.

A summary of the results are shown in Figure 1. The greatest difference was found for topic 1, which consists of the process, environment, and economic topics. The Colorado Springs Gazette was found to use this grouping of topics more often than the other two newspapers on average. This is likely due to the Colorado Springs Gazette using more technical words, along with a greater propensity to use economic words, than the other newspapers. Another example of the differences is for topic 2, which consists of moratorium, fracking, and regulation, and this group of topics is more prevalent in use in the Boulder Daily Camera. Topic 8, which consists of economic, regulation, and rights had a higher proportion on average for the Colorado Springs Gazette, which is in line with the potential systematic differences in reporting by a conservative leaning newspaper.

The trends remain generally consistent when we examine the difference between news articles and editorials or perspective articles, as shown in Figures 2 and 3.

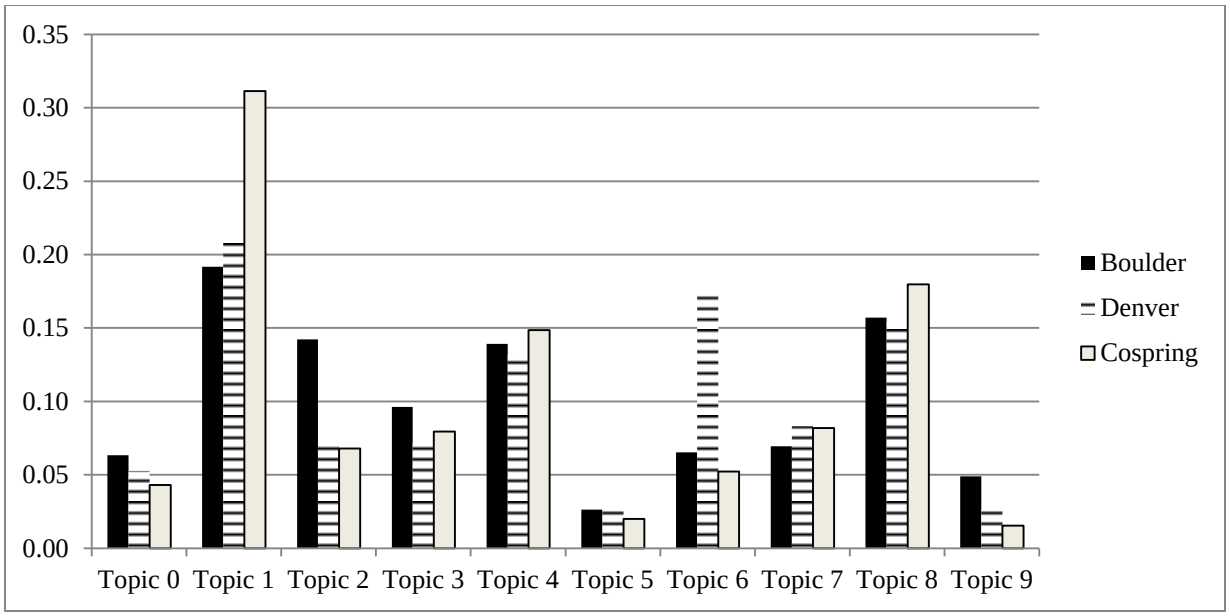

Figure 2. Average proportion of ten topics for news articles for the Boulder Daily Camera, Denver Post, and Colorado Springs Gazette. Topics are summarized in Table 7.

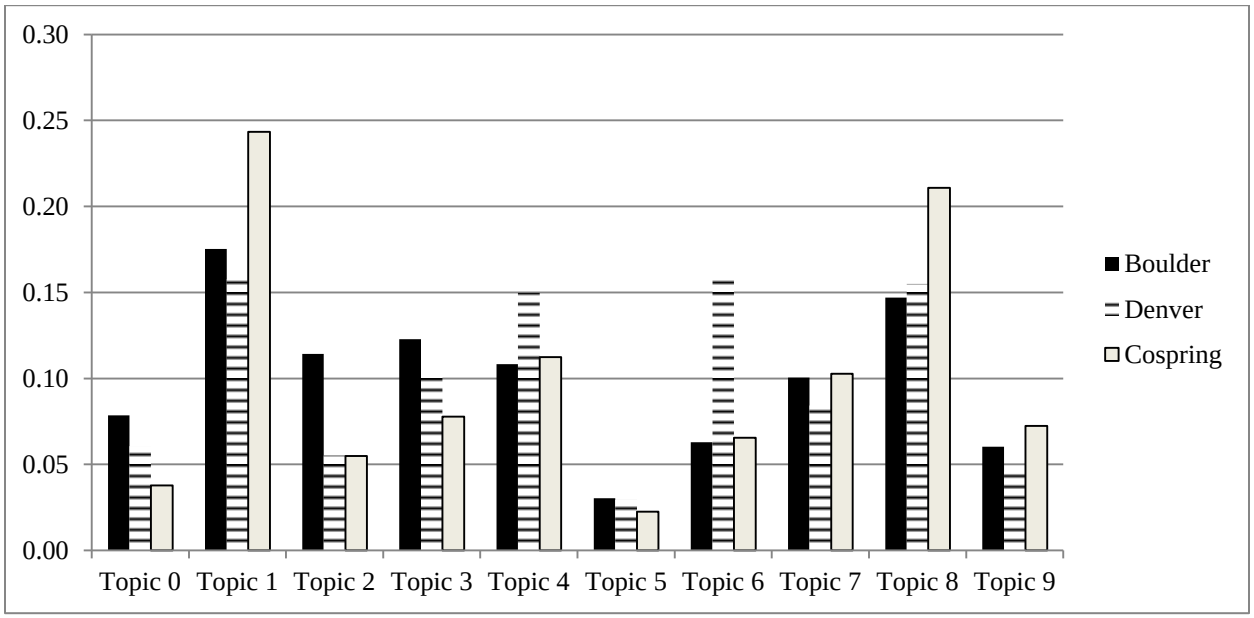

Figure 3. Average proportion of ten topics for editorials and opinion articles for the Boulder Daily Camera, Denver Post, and Colorado Springs Gazette. Topics are summarized in Table 7.

To further examine the differences between the Boulder Daily Camera and the Colorado Springs Gazette results from Figure 2 and 3, we used OLS to model the coefficient for each newspaper. The dependent variable is the proportion for each article. The independent variables are the dates of publication for the articles and a dummy variable to represent either Boulder or Colorado Springs. Tables 8 displays the unstandardized coefficient results for the news articles (editorials not included) and Table 9 displays the unstandardized coefficient results for the editorials. Significant results ($p<0.05$) are displayed in bold.

Table 8. OLS for News Articles Only (No editorials, letters, or perspectives).

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
R^2	0.01	0.08	0.13	0.01	0.01	0.01	0.25	0.01	0.00	0.03
F (df=3)	1.643	**13.915**	**24.326**	1.890	1.322	1.228	**53.973**	1.944	.745	**4.909**
					Unstandardized Coefficient					
	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
Constant	.280	1.179	**-2.875**	-.223	-.223	-.325	**2.572**	**.880**	-.103	.266
Date	-5.556E-06	-2.364E-05	**7.185E-05**	7.175E-06	7.175E-06	8.531E-06	**-5.849E-05**	-1.944E-05	6.270E-06	-5.861E-06
Boulder	.013	-.010	**.048**	.022	.022	-.002	**-.089**	-.007	.001	**.025**
CoSprings	-.008	**.104**	-.011	.008	.008	-.007	**-.115**	.001	.025	-.011

Table 9. OLS for Editorial, Letters, and Perspectives Only (No News Articles)

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
R^2	0.02	0.07	0.09	0.02	0.05	0.00	0.17	0.01	0.02	0.02
F (df=3)	3.606	**12.711**	**18.829**	**3.177**	**9.464**	0.410	**37.825**	0.968	3.717	**3.095**
					Unstandardized Coefficient					
	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
Constant	.349	**3.012**	**-2.125**	1.046	**-1.535**	-.099	**1.995**	.248	-.835	**-1.090**
Date	-7.034E-06	**-6.954E-05**	**5.31E-05**	**-2.295E-05**	**4.106E-05**	3.104E-06	**-4.478E-05**	-3.988E-06	2.414E-05	**2.764E-05**
Boulder	**.019**	**.033**	**0.047**	**.024**	**-.050**	.001	**-.084**	.017	-.014	.010
CoSprings	-.022	**.090**	-0.005	-.025	-.040	-.008	**-.090**	.018	.054	.026

Note: Coefficients are unstandardized coefficients for Table 8 and 9. Significant associations are listed in bold at $p<0.05$.

The differences between coefficients are larger between dimension and more are statistically significant for the editorial articles when compared to the news articles. This is partially expected due to the more sensational language often used in editorials and perspectives. Likewise, news articles are generally more balanced and less sensational. However, differences can be detected between the Boulder Daily Camera and Colorado Springs Gazette for both the news articles and editorials, after accounting for time, as shown in Tables 8 and 9.

Overall, there are differences across these three newspapers but not as substantial as expected. The strongest conclusion, at least based on the LDA, is that the three major newspapers in Colorado have more similar than dissimilar coverage on this controversial topic for both editorials and news articles.

3.4. Limitations

The LDA is most useful for finding the primary topic of an article and it is limited in its' ability to determine the varying subtopics within similar documents. In addition, model fitting and validation are content and source specific. Therefore, a substantial set of articles needs to be human coded for model fitting and then validation (Hopkins and King, 2010). Furthermore, an issue with this type of assessment is that the high proportion only detected the co-occurrence of a dimension or a topic, not the salience or context of these topics.

A significant challenge is the conversion of the proportion from the LDA output to a binary value that represents whether a topic is present. In this work, we fitted the model using a cutoff value for each topic, but this value can vary greatly and it can be influenced by the presence or absence of other topics. This approach is needed, for example, due to the potential of each topic having a value of 0.33 and this could be interpreted as all three topics are equally present or all three topics are not present.

Likewise, the proportion of a dimension or a topic does not necessarily imply the presence of a topic. Many articles quickly mention the topic, but it may not be the main focus of the article. Therefore, the proportion of that category is low, even though the topic is captured during human coding. In addition, human coding could utilize frequency coding, rather than simply the absence or presence of a theme, which may offer better reliability from a method like the LDA.

4.0. Conclusion

We found that automated coding through the use of modified LDA techniques can provide significant insight into the content of complex documents. Although automated coding is not a replacement for human coding, it does offer many opportunities. First, it can be used for a mile high view of the content and potential topics within a corpus. This could be used for experimental design prior to the start of a human coding project. Likewise, it could offer insight into the use of various documents that could eliminate less than ideal documents prior to starting human coding projects. At this time, automated methods are not a substitute for careful thought and close reading and require extensive and problem-specific validation (Grimmer and Stewart 2013). In addition, as stated by Grimmer and Stewart (2013) "for automated text methods to become a standard tool for political scientists, methodologists must contribute new methods and new methods of validation".

This report presents results of an automated media analysis of articles involving oil and gas development that utilizes hydraulic fracturing in Colorado. Below we summarize the key findings according to each of the primary study objectives.

1. *To compare the results of human and automated coding of media text.* Comparing the automated coding results to the human coding results, the inter-rater reliability was considered "fair". When the news articles that lacked a human coded topic were removed from the analysis, the percent agreement increased to 64%. The results also showed the accuracy for environment and economic topics to be 65% and 67%, respectively, which is comparable to automated coding results from other studies. This analysis suggests that automated coding using LDA can be used to understand the propensity of different topics across a large number of articles at a broad level.

2. *To conduct automated media analysis of liberal and conservative newspapers to assess differences in coverage on the hydraulic fracturing issue.* We found that the more conservative leaning newspapers use a greater proportion of economic words, while the liberal newspapers use a greater proportion of public health and environmental damage words. Using the Automap method LDA, we found that significant differences can be detected in the co-occurrence of topics based on the different newspapers evaluated.

In summary, these results demonstrated the ability to utilize automated coding methods to analysis the news media for the potential presence or absence of topics, along with the co-occurrence of topics within articles. Despite the limitations of automated coding, we hope to offer interested individuals and organizations a better understanding of the media framing of one of the most politically contentious environmental issues today in Colorado.

5.0. Works Cited

Adgate, J. L., Goldstein, B. D., & McKenzie, L. M. (2014). Potential public health hazards, exposures and health effects from unconventional natural gas development. Environmental science & technology.

Associated Press, Available through the Denver Post. (2012). Denver Post chooses Obama for 2nd term. Available Online at http://www.denverpost.com/politics/ci_21810641/denver-post-chooses-obama-2nd-term

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3, 993-1022.

Boulder Daily Camera. (2012). Endorsements 2012. Available online at http://www.dailycamera.com/ci_21916167/endorsements-2012

Boykoff, M. T., & Boykoff, J. M. (2004). Balance as bias: global warming and the US prestige press. Global environmental change, 14(2), 125-136.

Brown, E., Hartman, K., Borick, C. P., Rabe, B. G., & Ivacko, T. M. (2013). The National Surveys on Energy and Environment Public Opinion on Fracking: Perspectives from Michigan and Pennsylvania. Center for Local, State, and Urban Policy (CLOSUP), 'Survery Report: Climate Policy Options.

Carley, K. M., Columbus, D., & Azoulay, A. (2012). AutoMap User's Guide 2012(No. CMU-ISR-12-106). Carnegie – Mellon University, Pittsburgh, PA, Institute of Software Research International.

Carvalho, A. (2007). Ideological cultures and media discourses on scientific knowledge: re-reading news on climate change. Public understanding of science,16(2), 223-243.

Engelder, T. (2011). Natural gas: Should fracking stop?. Nature, 477(7364), 271-275.

Evensen D, Clarke C, Stedman R. 2014. A New York or Pennsylvania state of mind: social representations of gas development in the Marcellus Shale. Journal of Environmental Studies and Sciences, 4, 65-77.

Gallup Poll. (2013). In U.S., Trust in Media Recovers Slightly From All-Time Low. Available at: http://www.gallup.com/poll/164459/trust-media-recovers-slightly-time-low.aspx

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis.

Heikkila,T., Weible, C.M., Pierce, J., Gallaher, S. Kagan, J., Blair, B. (2013). A Summary Report of the Politics of Shale Gas Development and High-Volume Hydraulic Fracturing in New York. School of Public Affairs University of Colorado Denver.

Heikkila, T., Pierce, J. J., Gallaher, S., Kagan, J., Crow, D. A., & Weible, C. M. (2014). Understanding a Period of Policy Change: The Case of Hydraulic Fracturing Disclosure Policy in Colorado. Review of Policy Research, 31(2), 65-87.

Howarth, R. W., Ingraffea, A. (2011).Natural gas: Should fracking stop?. Nature, 477(7364), 271-275.

Jacquet J. 2014. Review of risks to communities from shale energy development. *Environmental Science & Technology* 48: 8321-8333.

Kinnaman T. 2011. The economic impact of shale gas extraction: A review of existing studies. Ecological Economics 70: 1243-1249.

McCallum, A. K. MALLET: A Machine Learning for Language Toolkit. 2002 [Htt p://mallet. cs. umass. edu].

McKenzie, L. M., Witter, R. Z., Newman, L. S., & Adgate, J. L. (2012). Human health risk assessment of air emissions from development of unconventional natural gas resources. Science of the Total Environment, 424, 79-87.

Macnamara, J. (2005). Media content analysis: Its uses, benefits and best practice methodology. Asia Pacific Public Relations Journal, 6(1), 1-34.

New York Times. (2012). Newspaper Endorsements From Across the Nation. Available online at: http://www.nytimes.com/interactive/us/politics/newspaper-presidential-endorsements.html?_r=0

Osborn, S. G., Vengosh, A., Warner, N. R., & Jackson, R. B. (2011). Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. proceedings of the National Academy of Sciences, 108(20), 8172-8176.

Pierce, J., Kagan, J., Heikkila,T., Weible, C.M., and Gallaher, S. (2013). A Summary Report of Perceptions of the Politics and Regulation of Hydraulic Fracturing in Colorado. School of Public Affairs University of Colorado Denver.

Porter, M. F. (1980). An algorithm for suffix stripping. Program: electronic library and information systems, 14(3), 130-137.

Shanahan, E. A., McBeth, M. K., & Hathaway, P. L. (2011). Narrative policy framework: The influence of media policy narratives on public opinion. Politics & Policy, 39(3), 373-400.

Shonkoff, S. B., Hays, J., & Finkel, M. L. (2014). Environmental Public Health Dimensions of Shale and Tight Gas Development. Environ Health Perspect.

Stedman, R. C., Jacquet, J. B., Filteau, M. R., Willits, F. K., Brasier, K. J., & McLaughlin, D. K. (2012). Environmental reviews and case studies: Marcellus shale gas development and new boomtown research: Views of New York and Pennsylvania residents. Environmental Practice, 14(04), 382-393.

Warner, N. R., Christie, C. A., Jackson, R. B., & Vengosh, A. (2013). Impacts of shale gas wastewater disposal on water quality in western Pennsylvania.Environmental science & technology, 47(20), 11849-11857.

Washington Post, 2004. 2004 Presidential Endorsements. Available online at: http://www.washingtonpost.com/wp-dyn/articles/A40874-2004Oct17_3.html

Wilson, K. M. (1995). Mass media as sources of global warming knowledge. Mass Comm Review, 22, 75-89.

Van Holt, T., Johnson, J. C., Brinkley, J. D., Carley, K. M., & Caspersen, J. (2012). Structure of ethnic violence in Sudan: a semi-automated network analysis of online news (2003–2010). Computational and Mathematical Organization Theory, 18(3), 340-355.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360-363.

Appendix A: Dictionary Method LDA Code

Example Mallet Code:

bin\mallet import-dir --input c:\mallet\mydata\cosprings --output train.mallet --keep-sequence --remove-stopwords

bin\mallet train-topics --input train.mallet --num-topics 2 --num-top-words 30 --num-iterations 100 --inferencer-filename inferencer --output-doc-topics doc-topics.txt --output-topic-keys cosprings.txt --output-doc-topics cosprings_compostion.txt

Bin\mallet import-dir --input test/ --keep-sequence --remove-stopwords --output test.mallet --use-pipe-from Train.mallet

Bin\mallet infer-topics --input test.mallet --inferencer inferencer --output-doc-topics incoming.topics1 --num-iterations 10 --output-doc-topics cosprings_compostion2.txt

Appendix B: Automap Method LDA Code

Example Mallet Code:

bin\mallet import-dir --input c:\mallet\mydata\cosprings --output cosprings.mallet --keep-sequence --remove-stopwords

bin\mallet train-topics --input cosprings.mallet --num-topics 15 --optimize-interval 15 --num-top-words 4 --output-state topic-cosprings.gz --output-topic-keys cosprings.txt --output-doc-topics cosprings_compostion.txt