HARVESTING WORDS

Efficient Dictionary Construction Techniques for Big Media Analysis

Kohei Watanabe

April 14, 2016

In recent years, computerized content analysis has been increasingly popular in social research. The forefront of this computerization seems to be political science, where new statistical techniques have been adopted from computer science or created for analysis of legislative speech transcripts or political party manifestos (e.g. Benoit & Laver, 2003; Lauderdale & Herzog, 2014; Slapin & Proksch, 2008). Political scientists owe their lead in computerized content analysis much to the public access to political documents in electronic formats. Yet, it is not only political documents that have become available in electronic formats in recent years; newspaper articles or TV news transcripts are also available on the internet or in commercial databases (e.g., Nexis or Factiva). The increased availability of electronic media data could encourage adoption of computerized content analysis in media research, popularising large-scale analyses of news content, but computerization in media analysis is still very limited.

The slow adoption of computerized content analysis to media research can be explained by a numbers of factors, which include copyright protections of media data and a weak quantitative tradition of the field, but one of the reasons might be a challenge researchers face in analysing media content with computer programs. The challenge stems from the greater diversity in content of news articles, which discuss a wider range of topics and opinions than elitist political documents; the diversity of media content even increases when their analyses span over a long period of time. The diverse topics and opinions in news stories results in larger variety of words, making statistical inference of meaning of words less reliable as they occur infrequently in the documents. Reliability of statistical inference could be improved by adding

more data to fit models, but this demands a large number of news articles to be manually content analysed just in preparation for content analysis, diminish the very advantage of the computerization.

Nonetheless, there were studies, in which researchers successfully performed computerized content analysis of news. For example, Roberts and McCombs (1994) identified media agenda in newspaper articles; Kellstedt (2000) analysed framing of racial issues in *Newsweek* in terms of egalitarianism or individualism. More recently, Segev and Miesch (2011) identified framing on the Israeli-Palestinian conflict in news in six languages; Young and Soroka (2012) predicated opinion poll results by analysing news coverage of a Canadian federal election. In all these studies, content analysis was based on a pre-defined list of words, called 'dictionary'. The key to their success were accurate classification of words by researchers themselves based on their expert knowledge on the subjects.

However, construction of original dictionaries requires a large amount of time. Creation of original dictionaries would be avoided by adopting existing dictionaries such as the General Inquirer Dictionary (Stone, Dunphy, Smith, & Ogilvie, 1966), LIWC (Francis & Pennebaker, 1993), the Regressive Imagery Dictionary (Martindale, 1975) and DICTION (North, Lagerstrom, & Mitchell, 1984), but adoption of these off-the-self dictionaries raises concerns regarding validity of measurements due to the lack of transparency of the dictionary making procedure (Grimmer & Stewart, 2013b; Neuendorf, 2002). Validation of results of computerized content analysis would dismiss such concerns, but existing dictionaries often fail to pass validation tests when they are applied to documents in different subjects from those they were originally developed for, because words, and their meanings, in documents vary subject to subject.

The purpose of this paper is presenting a new content analysis technique called Latent Semantic Scaling (LSS) developed by the author, aiming to solve the above-mentioned problems in computerized analysis of media content. LSS constructs content analysis dictionaries very efficiently with minimal human involvement, exploiting rich information in a large corpus of news articles for reliable estimation of

semantic values of words. The basic application of LSS is construction of subject-specific sentiment dictionaries, but it can also be applied for construction of dictionaries that measure more complex dimensions by selecting a small set of exemplary words called ‘seed words’. Further, even when users are unable to select seed words, LSS constructs dictionaries that replicate manual content analysis through automated selection of seed words as a supervised machine learning technique. The dictionaries constructed by LSS in either way can be applied to content analyse large number of news stories fully automatically.

This type of computerized content analysis has been little explored in social sciences, but recognized as ‘lexicon expansion’ in computer science, where many different techniques have been developed (c.f. Liu & Hu, 2004). Among those, Sumbaly and Sinha (2009) successfully extracted a sentiment words from a corpus of *New York Times* articles, but LSS is different in an important respect from existing techniques. Whereas existing lexicon expansion techniques were based on either co-occurrence (collocation) or spatial proximities (LSA) of words (Landauer & Dutnais, 1997; Turney & Littman, 2003), LSS utilizes both collocation and spatial analyses, respectively, for selection of feature words related to subjects and estimation of semantic values of the feature words. This combination of the two different of analyses makes LSS dictionaries robust against irrelevant words in news stories, while maintaining high sensitivity to subtle differences in news content.

My discussion on this dictionary construction technique in this paper is separated into two parts. In the first part, I will explain the basic LSS technique by constructing highly domain specific sentiment dictionaries, which gauge positive-negative tones of news stories on democracy or sovereignty in Ukraine. These dictionaries were constructed for my own research project on Russian news agencies’ framing of the Ukraine crisis in their English-language services (Author forthcoming). The accuracy of content analysis by the dictionaries will be then compared with a off-the-self sentiment dictionary constructed by Young and Soroka (2012). In the second part, I will explain the supervised LSS technique by constructing a dictionary on Russian state-controlled media’s framing of street protests (Lankina et al.

forthcoming). This dictionary measures more conceptually complex than positive-negative sentiments: whether street protests were framed as freedom of expression or public disorder by the media. Its accuracy will be compared with Wordscore (Benoit & Laver, 2003), which is an example of supervised Bayesian technique, to show core advantages of LSS in analysis of media content. This example also demonstrates the ability of the LSS to handle non-English languages.

Part 1: Growing seeds

Construction of subject-specific sentiment dictionaries by LSS requires almost no human involvement, because the computer program automatically selects and scores words relevant to a given subject. In this technique, automated construction of dictionaries is achieved by combining (a) collocation analysis for feature selection and (b) Latent Semantic Analysis (LSA) for

sentiment estimation based on statistical analyses of a large news corpus. I will explain the details through a construction of dictionaries that gauge positive-negative tones in news stories on democracy or sovereignty published by Russian news during the Ukraine crisis.

Sentiment dictionaries on Ukraine crisis

For this example, I constructed a large corpus of English-language news stories published by ITAR-TASS, Interfax, and Reuters between January 2013 and December 2014, downloading 240,173 full-text articles from news databases. In preparation for dictionary making, I split all the news articles into sentences, and removed all the proper names. This pre-processing is essential for LSS to accurately estimate semantic relationship between words.[1] I will not discuss methods to identify proper nouns is news articles, but one can employ any effective tools for this purposes.

[1] Units has to be sentences because immediate contexts of words are important to make inference on their meanings; proper nouns as well as proper adjectives have to be removed from the corpus because they skew general meanings of words.

Selecting relevant words

In order to select features that are strongly associated with democracy or sovereignty, I analysed frequencies words co-occur with ‘democracy’ or ‘sovereignty’ in the corpus. This type of analysis, called collocation analysis, is a common method to statistically estimate association between different words in a corpus. In this case, collocations were any words occurring within 10 words from the target words that were defined by patterns (‘democra*’ or ‘sovereign*’) in the corpus. The relevance of the collocations were measured by the likelihood ratio statistic, or G-score (Hoey, 2012). To computed G-score, I count the occurrence of a word w_i and all other words $\overline{w}_i$ within 10 words ($d_i \leq 10$) from the target words, and construct contingency tables:

	$d_i \leq 10$	$d_i > 10$
w_i	n_1	n_2
$\overline{w}_i$	n_3	n_4

With such contingency tables, I obtained G-scores g_i for w_i by comparing observed counts n_j with expected counts e_j, which were estimate by the marginal destitution of the observed counts in the same way as chi-square test:

$$g_i = 2\sum_{j}^{J=4} n_j \cdot \log\left(\frac{n_j}{e_j}\right) \tag{1.1}$$

I selected up to 1,000 features with $g_i > 10.83$, the critical value for 99.9% confidence level, provided that their observed counts were greater than their expected counts, $n_1 > e_1$. The number of features selected by these criteria was 778 for democracy and 626 for sovereignty. All features in Table 1 for democracy seem to be intuitively related to the topic, but there are a few financial words such as ‘debt’ and ‘bonds’ for sovereignly. These financial words seem to be erroneous selected, but they have limited impact on final outcomes of content analysis.

Table 1: Top 20 features for democracy and sovereignty

Rank	Democracy	G-Score	Sovereignty	G-Score
1	human-rights	8,503.6	integrity	11,446.2
2	institutions	3,292.0	territorial	8,216.5
3	law	3,055.1	independence	2,932.5
4	supremacy	2,938.1	respect	2,097.4
5	elections	2,649.4	state	2,043.2
6	reforms	2,505.5	rating	1,681.2
7	rule	2,494.0	states	1,245.6
8	values	1,934.9	right	1,223.7
9	principles	1,572.3	debt	1,200.6
10	freedom	1,515.7	country	974.2
11	election	1,486.8	ratings	947.7
12	standards	1,468.9	independent	880.6
13	freedoms	1,161.3	bonds	837.1
14	party	1,151.6	dispute	698.0
15	society	1,116.5	islands	661.5
16	country	1,051.3	principles	656.9
17	political	911.2	non-interference	536.2
18	free	859.3	national	497.0
19	opposition	793.3	internal	493.1
20	parliamentary	727.8	violation	464.0

Assigning sentiment scores to words

For the 778 and 626 features entered into dictionaries, I estimated of sentiment scores by analysing their proximity to the general English positive and negative ‘seed words’, which were originally identified by Turney and Littman (2003) and have been widely used in computer science literature. The seed words are {good, nice, excellent, positive, fortunate, correct, superior} and {bad, nasty, poor, negative, unfortunate, wrong, inferior}.

The simplest approach to estimation of semantic proximity between words is calculating cosine similarity of row vectors of term-document matrix, but my term-sentence matrix X with 270,000 rows and over million columns, which I obtained from the large corpus, was too sparse for estimation of semantic proximity (left in Figure 1). Therefore, I decompose the matrix to three matrices, U, D and V, utilizing Singular Value Decomposition (SVD), and constructed a matrix $\hat{S}$ with only 300 columns (right in Figure 1). This technique is known as Latent Semantic Analysis (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990), and can be understood as Principal Component Analysis of term-document matrix.

$$X \approx \hat{X} = UDV' \tag{1.2}$$

$$\hat{S} = UD \tag{1.3}$$

With the matrix $\hat{S}$, I estimated sentiment of words by their average cosine similarity to the seed words: The sentiment score v_i for a word w_i was a mean cosine similarity to seed words weighted by seed scores p_j, which were simply +1 for the positive seeds words and −1 for the negative seeds words. Here $\cos(w_i, s_j)$ denotes cosine similarity between two row vectors corresponding to word w_i and s_j in the matrix $\hat{S}$.

$$v_i = \frac{1}{n}\sum_{j}^{n} \cos(w_i, s_j) \cdot p_j \tag{1.4}$$

Table 2 and 3 present top 20 most positive and negative words on democracy and sovereignty. In both tables, many of words are intuitively positive or negative, but some are not. For example, ‘intensify’ in democracy, is not always used in positive context, but we are not able to judge if its score is accurate, unless we investigate its usage in the large corpus. Also, ‘upon’ in sovereignly is a function word lacking substantive meaning, but it could be remove easily if larger list of stopwords would have been utilized.

Figure 1: Notional illustration of dimension reduction by SVD

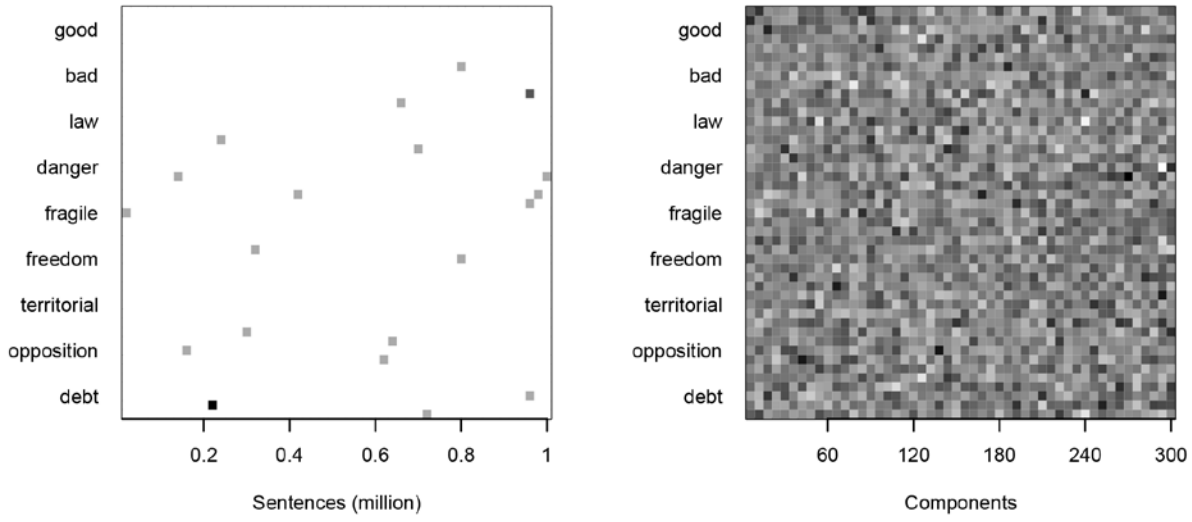


Table 2: Most positive and negative entry words for democracy

Rank	Positive	Sentiment Score	Negative	Sentiment Score
1	Normalising	0.0075	bears	-0.0142
2	inter-parliamentary	0.0074	danger	-0.0138
3	Tangible	0.0068	fear	-0.0129
4	praised	0.0066	threatening	-0.0126
5	intensify	0.0060	inability	-0.0121
6	strengthening	0.0058	pose	-0.0111
7	strengthen	0.0055	blow	-0.0110
8	establishment	0.0053	themselves	-0.0110
9	co-operation	0.0047	itself	-0.0106
10	peoples	0.0045	helping	-0.0105
11	milestone	0.0041	voice	-0.0105
12	consolidate	0.0040	moving	-0.0103
13	contribute	0.0040	strong	-0.0103
14	importance	0.0039	strongly	-0.0103
15	develop	0.0037	watchdog	-0.0103
16	dialogue	0.0037	criticism	-0.0101
17	achieve	0.0034	beacon	-0.0100
18	chairman	0.0034	fragile	-0.0099
19	promote	0.0034	posed	-0.0097
20	upcoming	0.0033	deeply	-0.0096

Table 3: Most positive and negative entry words for sovereignty

Rank	Positive	Sentiment Score	Negative	Sentiment Score
1	relations	0.0099	risk	-0.0166
2	allied	0.0069	low	-0.0137
3	normalise	0.0067	threatening	-0.0126
4	strengthening	0.0058	lose	-0.0124
5	mutual	0.0056	default	-0.0122
6	strengthen	0.0055	negative	-0.0120
7	establishment	0.0053	risks	-0.0114
8	positive	0.0047	likelihood	-0.0113
9	peoples	0.0045	wealth	-0.0112
10	unquestionable	0.0042	pose	-0.0111
11	thanked	0.0040	themselves	-0.0110
12	upon	0.0040	survival	-0.0107
13	contribute	0.0040	itself	-0.0106
14	importance	0.0039	consequence	-0.0106
15	develop	0.0037	threatened	-0.0105
16	dialogue	0.0037	loss	-0.0104
17	adherence	0.0033	afford	-0.0104
18	invariable	0.0031	strong	-0.0103
19	pragmatism	0.0030	strongly	-0.0103
20	commitment	0.0029	posing	-0.0103

Once sentiment scores were assigned to entry words, the dictionaries became ready for content analysing news articles. When entry words $w_{i...I}$ occur in a story in total of m times, and v_i is the sentiment score and f_i is the frequency count of an entry word w_i, its document score d' is computed as:

$$\acute{d} = \frac{1}{m}\sum_{i}^{I} v_i \cdot f_i \tag{1.5}$$

Comparing LSS and LSD

I applied the LSS dictionaries to two samples of news stories on democracy or sovereignty of Ukraine published by the Russian agencies (ITAR-TASS and Interfax) to test its validity. I also applied Lexicoder Sentiment Dictionary (LSD) as an example of off-the-shelf dictionary to the samples.[2] In dictionary-based content analysis tools including Lexicoder, the sentiment score of a document d is the difference between normalized frequency of positive or negative words, which is defined as:

$$d = \frac{n_{\text{pos}} - n_{\text{neg}}}{l} \tag{1.6}$$

where n_{pos} and n_{neg} are numbers of positive or negative words in a dictionary, and l is the total number of words in the document.

Figure 2 and 3 compare document scores assigned by computerized content analysis (LSD or LSS) with those by manual content analysis.[3] In stories on democracy, scores computed by LSD correlated with human scores in many of the cases, but overall correlation is only moderate (r=0.46) due to the overestimation of positivity (#6, #16) or negativity (#26). In LSS, there are two large errors (#8, #27), but

[2] Lexicoder has sophisticated negation words handling mechanism, but I adopted the common bag-of-words approach.

[3] Very small raw scores of LSS are recalled to between −100 and +100. In manual coding, I classified on five-point scale {1: very negative, 2: negative, 3: neutral, 4: positive, 5: very positive}, and calculated document score by taking means of sentence scores.

others cases were accurately scored, achieving, stronger correlated with human scores (r=0.77). In sovereignty, however, LSD underestimated positivity of many of the stories, but extreme stories (#6, #21, #25) were scored very accurately, hugely affecting the correlation coefficient (r=0.65). LSS was less accurate (r=0.70) in sovereignty than in democracy, creating random error, but outperformed LSD in this topic too.

Figure 2: Machine coding of stories on democracy

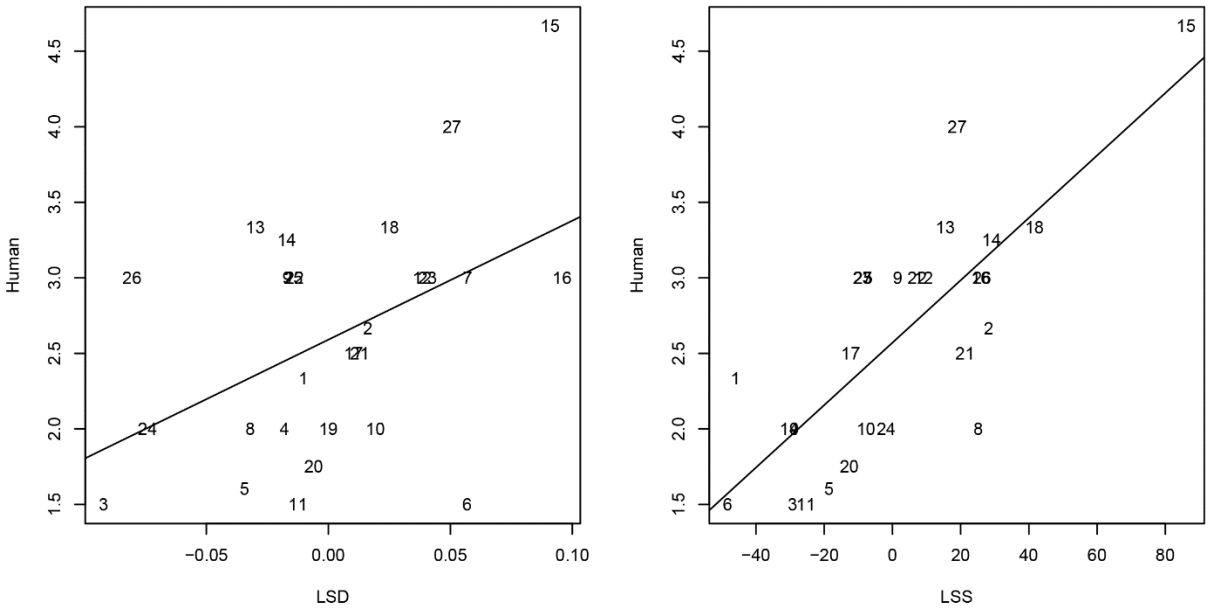
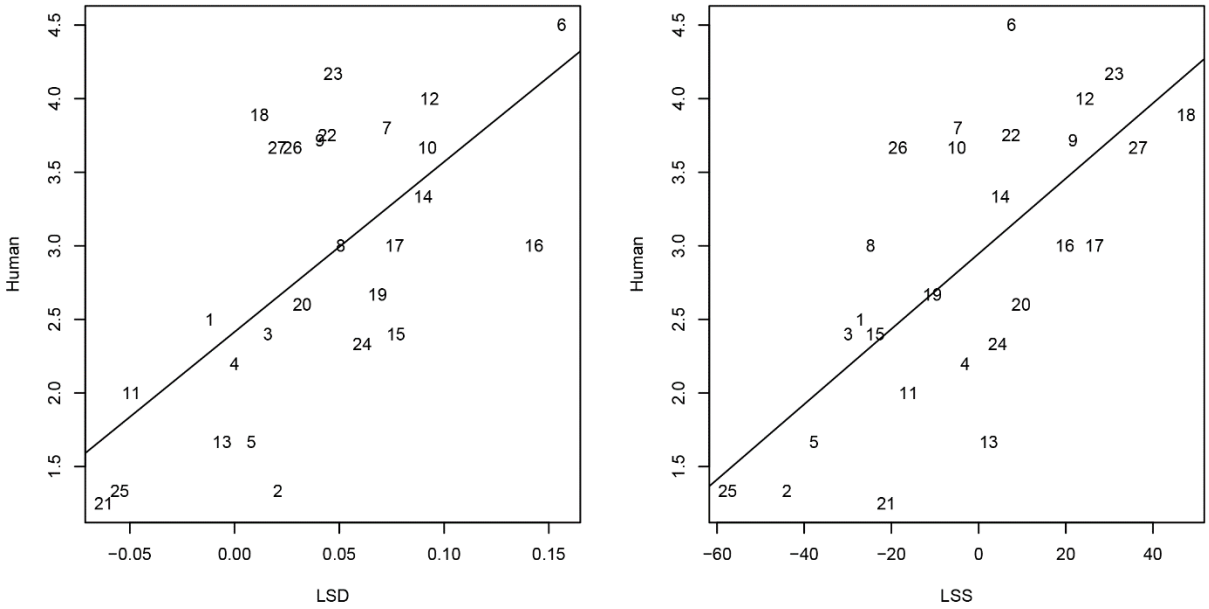


Figure 3: Machine coding of stories on sovereignty



Part 2: Mining seeds

I have constructed sentiment dictionaries from the existing seed words, but nothing prevents users from selecting and applying home-grown seed words to LSS. However, manual selection of seed words is sometimes very difficult to achieve, particularly when users wish to gauge very complex dimensions of news content. In this part of the paper, therefore, I will explain a technique that allows users to construct dictionaries on almost any given topics and dimensions without manually selecting seed words. This technique adds an automated seed selection process to LSS, but is still based the original LSS algorithm for selection and scoring of entry words.

Framing dictionary on Russian street protests

In this part, I will explain how I created a dictionary that measures how street protests in Russia were framed by Russian media using an automated seed words selection technique. The dimension is as freedom of expression vs. public disorder, a dimension much more complex than positive vs. negative. For the dictionary, I constructed a corpus of news stories published by state-controlled newspapers and TV broadcasters in Russia (Channel 1, NTV, Russia 1, *Izvestiya, Komsomolskaya Pravda*, and *Russian*

Gazette) between 2011 and 2014. The corpus contained 39,787 Russian-language full-text news articles or transcripts on street protests. Note that just like in the first part, I unitized news articles into sentences and eliminated all the proper nouns and proper adjectives from the documents before applying LSS.

I also took a random sample of 30 news articles from the corpus, and asked native Russian speakers to classify each sentence of the articles on 5-point scale ranging from 'explicitly as social disorder' to 'explicitly as freedom of expression'. I aggregated the sentences scores to obtain accurate document scores, and allocated the first half documents for dictionary making (training set), and the last half for validation (test set).

Finding seed words

The goal in my seed word selection was to identify 5 to 10 pairs of words that define the freedom-disorder dimension, but there were numerous potential seed words. Therefore, I restricted candidates for seed words to those strongly associated with "protest" ("протест*"), which were identified by collocation analysis (Equation 1.1), because I wished subject specific dictionary to be created from subject specific seed words. I selected top 10% of collocations, totalling 1,038 words with minimum G-score of 55.4. (Since this was the same procedure for feature selection, seed words were chosen from entry words to the dictionary in this case).

To test suitability of each seed candidate, I had to create a large number of tentative dictionaries, but it was completed very quickly by calculation pair-wise cosine similarities between all these seed candidates in the beginning. I calculated pair-wise cosine similarities in a SVD-reduced matrix $\hat{S}$ (Equation 1.3) that I created from the corpus. The cosine similarities for all pairs were stored in a symmetric matrix D, which has $K = 1{,}038$ columns and rows corresponding to the seed candidates $c_{k\ldots K}$. Given the similarity matrix D, a temporary dictionary for a seed word c_k is a kth row or column vector of the matrix D.

$$d_k = D_{\cdot k} = D_{k\cdot} \tag{2.1}$$

First, I created 1,038 temporary dictionaries in this way, and applied them to the training set (Equation 1.5) to obtain correlation coefficients r_k between scores computed by the temporaries d_k and scores assigned by human coders. These correlation coefficients allowed me to infer importance and polarity of the seed candidates. The importance of seed candidates was measured by the sizes of the correlation coefficients; the polarity of seed candidates was by the signs of the correlation coefficients. I selected only 50 seed candidates with the largest absolute correlation coefficient from both sides of polarity, and assigned seed scores p_k in the following manner:

$$p_k = \begin{cases} +1, & r_k > 0 \\ -1, & r_k < 0 \end{cases} \tag{2.2}$$

Then, seed words were given adjusted scores to make scoring of documents more consistent when they are combined into single seed set. An adjusted seed scores $\acute{p}_k$was a seed score weighted by inverse of average squared similarity to other seed candidates in the matrix D (Equation 2.1):

$$\acute{p}_k = p_k \cdot \frac{1}{\sum D_{\cdot k}^2 \cdot \frac{1}{K}} \tag{2.3}$$

Second, with 100 seed candidates from the both sides, I constructed pairs of seed words $\{c_k, c_l\}$, searching for partner c_l for c_k such that (1) the partner has opposite polarity $p_l \neq p_k$, (2) the dictionary $d_{\{k,l\}}$ yields a higher correlation coefficient than the separate dictionaries $r_{\{k,l\}} > r_k$ AND $r_{\{k,l\}} > r_l$, and (3) the correlation become the strongest with the partner $r_{\{k,l\}} \geq r_{\{k,\bar{k}\}}$. Starting from the seed candidate with the larges absolute correlation coefficient $|r_k|$, all other seed candidates entered this step-wise paring process. This process continued until at least five pairs were found, and new pairs start decreasing the overall correlation. This process only took around 30 seconds on my laptop computer.

In the above process, I could easily construct a dictionary with $K = 1{,}038$ entry words with any set of seed words. Scores assigned to entry words $v_{k \ldots K}$ were calculated simply by taking inner products of and

the weighted seed scores and a subset of the similarity matrix $\acute{D}$ that only has columns corresponding to the seed words:

$$v_k = \acute{D} \cdot \acute{p}_k \tag{2.4}$$

Comparing LSS and Wordscore

The dictionary that I crated with the automatically chosen seed words could be applied to the 15 manually scored news articles (test set) in exactly the same way as those words produced with manually chosen seed words (Equation 1.5). I also applied a Bayesian document scaling technique Wordscore (Benoit & Laver, 2003) to the documents for comparison. In Wordscore, when there are H manually scored documents, the scores for a word v_i is its average frequency weighted by document scores d_h:

$$v_i = \sum_{h}^{H} \frac{f_i}{k_h} \cdot d_h \tag{2.5}$$

where k_h is the total number of words in the hth document. In training the Wordscore model, I eliminated words that do not occur more than five times to obtain the best result.

Table 4 shows Russian seed words which were automatically selected with the training set. Although it is difficult to judge the suitability of the seed words without knowledge regarding their contexts, the freedom-of-assembly seeds related to legal or administrative procedures, while the social-disorder seeds were words often used to describe attributes or the behaviour of protesters.

Table 4: Automatically selected protest framing seed words

Seed word	Seed word (English translation)	Seed Score
подали	filed	74.7
сопровождалось	accompanied by	58.7
атаковала	attacked	53.8
бессрочной	termless	44.7
стычки	clashes	39.3
основополагающие	fundermental	-48.9
использования	utilization	-98.1
подчиняющиеся	obeying	-130.2
госпереворот	coup	-149.0

нежелания	unwillingness	-306.0

Figure 4 shows document scores assigned to the training set by Wordscore or LSS. Wordscore much better reproduced scores assigned by human coders (r=0.93) than LSS (r=0.85); 95% confidence intervals are also very small in Wordscore, indicating high confidence in estimated scores. However, when their model or dictionary were applied to the test set, LSS (r=0.76) appeared much better than Wordscore (r=0.39) (Figure 5).

Figure 4: Machine coding of stories on protests (training set)

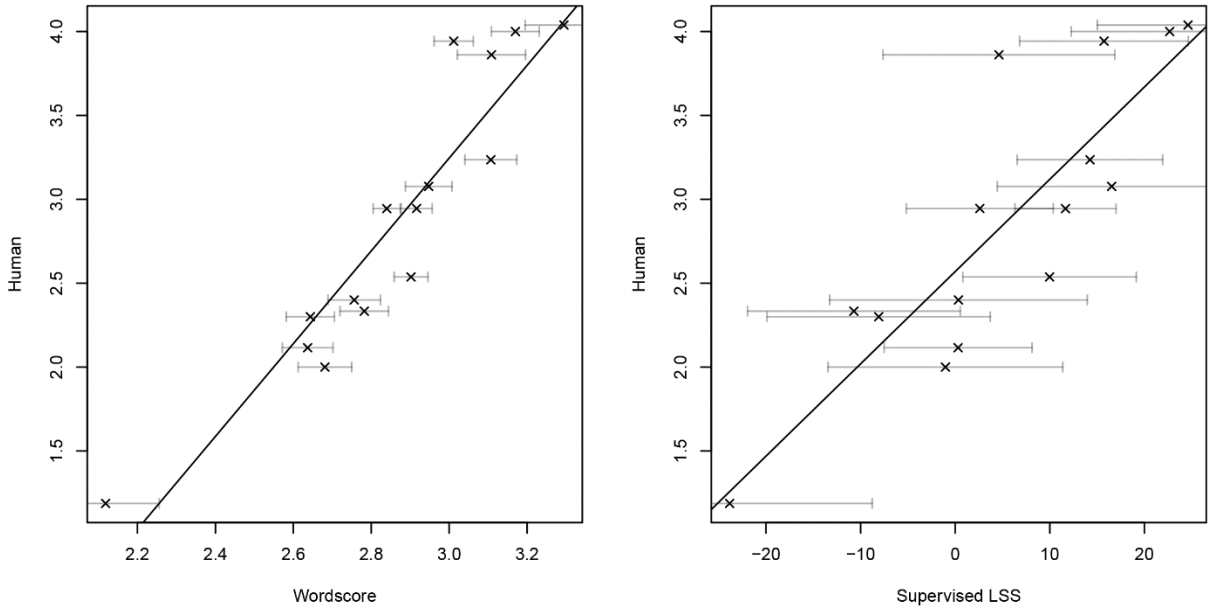
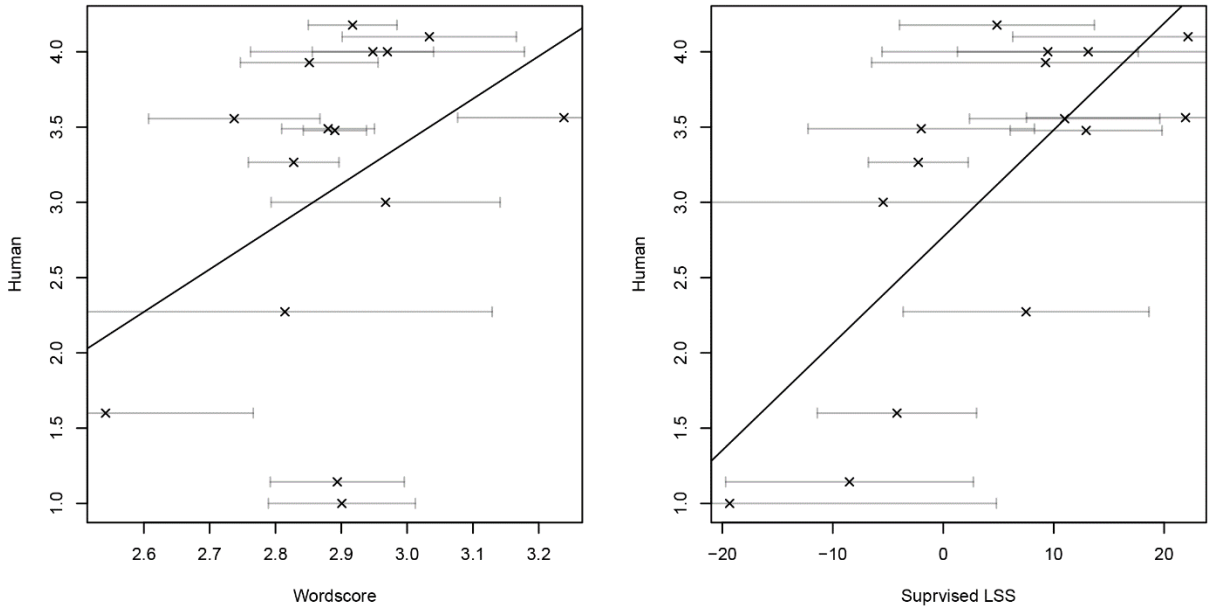


Figure 5: Machine coding of stories on protests (test set)



Discussion

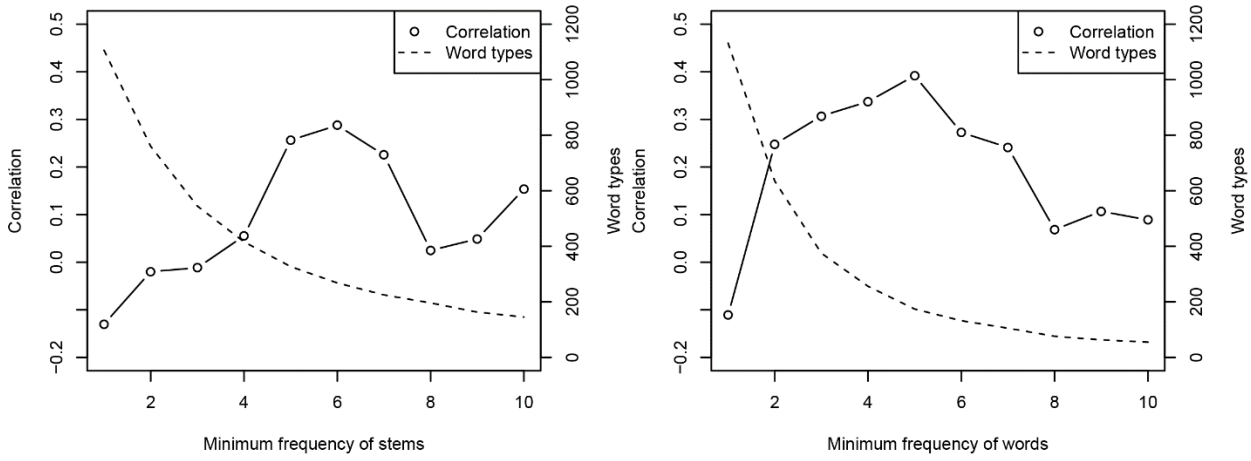
In the first part of this paper, I clearly showed that risk of adopting off-the-shelf dictionary without validation (Grimmer & Stewart, 2013a) as Lexicoder Sentiment Dictionary (LSD) only scored news on sovereignty sufficiently accurately. In this research, I did not utilize Lexicoder's sophisticated negation handing mechanism, which would increase accuracy of its analysis, but it does not seem to be the main cause of the problem. Instead, the reason seems to be that the words used by the Russian news agencies in reporting democracy in Ukraine were significantly different from those used by Canadian or other Western media report stories in their countries. This is exactly the problem I addressed by selecting and scoring words from the news corpus. The subject-specific dictionaries reflect the unique vocabulary and the style of writing of the Russia news agencies, and increased the accuracy in computerized content analysis.

However, I found a weakness of LSS sentiment dictionaries vis-à-vis the manually compiled dictionary. Namely, the LSS sentiment dictionaries were less accurate in scoring extremely positive or negative news articles than manually compiled. This seems to have been caused by the LSS algorithm that estimate

semantic values of entry words by their similarity to seed words. The general English positive seed words limited the range of sentiment scores within the relatively narrow semantic space that they define. As a result, words more positive or negative than the seed words were not scored correctly, LSS failing to recognize extremely sentiments in some of the stories. This property of LSS may require the seed set include extremely positive or negative words.

In the second part of this paper, a problem in adopting common Bayesian techniques to an analysis of news content became very clear. That is, Bayesian supervised learning algorithms only assign reliable score to words that occur frequently in training sets. In order to construct a best-performing Wordscore model (r=0.39), I had to eliminate words that do not occur more than five times. As shown in the right panel of Figure 6, when I included all the words in the training set, the model did not replicate human scoring at all (r=−0.11); when words only occur once in the training set were excluded, the correlation increase to r=0.24. In this way, an increase in the threshold for minimum frequency increased the correlation until the minimum frequency became five. The same trend was also found when stemming was performed (left in Figure 6).

Figure 6: Minimum frequency of tokens and token types



Nonetheless, the higher threshold for the minimum frequency of words did not solve the problem of the Bayesian method because it also decreased the number of word types appear in both the training set and

test set (broken lines in Figure 6). When the minimum frequency was one, there were 1,132 types of words, but the number halved when minimum frequency was increased to two. When the threshold was raised to five, it decreased to 175 types. After this point, the correlation started falling sharply as the model failed to recognize relevant words in the test set. This clearly shows the dilemma of Bayesian techniques that high frequency of words is necessary for reliable parameter estimation, but only a few words are frequent enough in small training sets. The only solution to this dilemma is to increase the size of training sets through expensive manual content analysis.

In contrast, LSS achieved to constructed a better-performing dictionary with 1,038 entry words, 200 of which appear in the test set, from the same small training set aided by the rich information in the large news corpus. All the entry words in the dictionary were selected and scored based on the information in the large news corpus, not directly from frequencies of words in the training set. Owing to the size of the corpus, the reliability of parameters, and therefore accuracy of analysis, were much higher in LSS than in the Bayesian method. In other words, construction of content analysis dictionaries by LSS does not depend on the size of training set but on the size of corpus, which we can easily increase at minimal costs by sourcing websites or databases.

Finally, I have to emphasize features of the LSS algorithm that allowed me to estimate of parameters for over a thousand words only from 15 documents in a training set: (a) the estimated semantic proximities of entry words, (b) the pre-prioritization of seed words, and (c) the assumed independence between seed words. Thanks to the estimated semantic proximities, I could estimate semantic values of 1,038 entry words by selecting only one word; prioritization of seed candidates reduced the potential partners to no more than 50; the assumed independence of seed words allowed me to combine all the pairs of seed words into the final seed set.

Conclusion

The benefit of LSS to the analysis of news content is not only the higher accuracy in scoring, but also efficiency. As I have shown by creating sentiment dictionaries in the first part, LSS demands almost no human involvement as far as large news corpus and seed words are available. Large news corpora can be constructed easily by downloading news articles from websites or databases. When users are unable to select seed word, supervised-LSS select seed words automatically. As I have demonstrated in the second part, the automated seed words discovery algorithm only requires only around 15 manually scored document. With this efficiency of LSS, media analysis will be empowered to embark on large-scale analysis of media content, expanding the horizon of the computerized content analysis.

Bibliography

Benoit, K., & Laver, M. (2003). Estimating Irish party policy positions using computer wordscoring: the 2002 election – a research note. *Irish Political Studies*, *18*(1), 97–107. http://doi.org/10.1080/07907180312331293249

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

Francis, M. E., & Pennebaker, J. W. (1993). *LIWC: Linguistic Inquiry and Word Count* (Technical Report). Dallas, Texas: Southern Methodist University.

Grimmer, J., & Stewart, B. M. (2013a). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political. *Political Analysis*. http://doi.org/10.1093/pan/mps028

Grimmer, J., & Stewart, B. M. (2013b). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 1–31. http://doi.org/10.1093/pan/mps028

Hoey, J. (2012). The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test. *arXiv:1206.4881 [Stat]*. Retrieved from http://arxiv.org/abs/1206.4881

Kellstedt, P. M. (2000). Media Framing and the Dynamics of Racial Policy Preferences. *American Journal of Political Science*, *44*(2), 245–260. http://doi.org/10.2307/2669308

Landauer, T. K., & Dutnais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 211–240.

Lauderdale, B. E., & Herzog, A. (2014). Measuring Political Positions from Legislative Debate Texts on Heterogenous Topics. Retrieved from http://www.alexherzog.net/files/Lauderdale_Herzog_2015.pdf

Liu, B., & Hu, M. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Seattle, Washington.

Martindale, C. (1975). *Romantic progression : the psychology of literary history*. Washington, DC: Hemisphere Publishing ; New York ; London.

Neuendorf, K. A. (2002). *Content Analysis Guidebook*. SAGE.

North, R., Lagerstrom, R., & Mitchell, W. (1984). *DICTION Computer Program: Version 1*. Retrieved from http://www.icpsr.umich.edu.gate2.library.lse.ac.uk/icpsrweb/ICPSR/studies/5909/version/1

Roberts, M., & McCombs, M. E. (1994). Agenda setting and political advertising: Origins of the news agenda. *Political Communication*, *11*(3), 249–262. http://doi.org/10.1080/10584609.1994.9963030

Segev, E., & Miesch, R. (2011). A Systematic Procedure for Detecting News Biases: The Case of Israel in European News Sites. *International Journal of Communication*, *5*(0), 20.

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722. http://doi.org/10.1111/j.1540-5907.2008.00338.x

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* The M. I. T. Press.

Sumbaly, R., & Sinha, S. (2009). Sentiment Mining in Large News Datasets. Retrieved from cs.stanford.edu/people/rsumbaly/files/Sentiment_Mining.pdf

Turney, P. D., & Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Trans. Inf. Syst.*, *21*(4), 315–346. http://doi.org/10.1145/944012.944013

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29*(2), 205–231. http://doi.org/10.1080/10584609.2012.671234