Big Media Analysis: Application of Vector Space Models to Document Scaling

Kohei Watanabe

London School of Economics and Political Science

February 20, 2017

5,778 words

Author's note

The LSS technique was originally developed in Python, but recently implemented in R. R package LSS is available at https://github.com/koheiw/LSS.

Abstract

Computerized analysis of media content is often challenging because diverse topics in news stories cause high data sparseness. Although supervised machine learning techniques usually requires large training set for accurately analysing diverse content, this paper proposes use of vectors space models for the purpose. Vectors space models, such as LSA, NMF, LDA or Word2vec, are used to extract semantic information from large corpora fully automatically to reliably estimate parameters for words that rarely appear in small training set. This new technique is explained with two examples from actual large-scale content analysis projects: international news agencies' coverage of the Ukraine crisis 2013-2014, and Russian news media's coverage of street protests 2011-2014. These examples show advantages of the new technique in document scaling over 'off-the-self' dictionaries and Bayesian supervised-machine learning techniques.

Computerized content analysis has become increasingly popular in social research. The forefront of this computerization seems to be political science, where new statistical techniques have been created for the 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 availability of 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 have also become available on the internet or in commercial databases (e.g., Yet, it is not only political documents that have become available in electronic formats in recent years; newspaper articles or TV news transcripts have also become available on the internet or in commercial databases (e.g. Nexis or Factiva).

However, computerization in content analysis is still very limited in media studies despite the increased availability of electronic data. Several factors explain the slow adoption of computerized content analysis for media research, including issues of copyright protection of the media sources and qualitative orientation of the field, but the greater content diversity of news articles explains more: content analysts face greater challenges in computerization, because news articles discuss a wider range of topics and views than political documents. The diverse topics and opinions in news stories result in a larger variety of vocabulary, thus most documents share only up to 1-5% of words with other documents (Phan, Nguyen, & Horiguchi, 2008). Add-one smoothing and word stemming have been commonly used to address this data sparseness problem in computerized content analysis (Manning, Raghavan, & Schütze, 2008). All the words are given the pseudo count of 1 regardless of the patterns of their occurrences in the data in add-one smoothing, or words endings are mechanically truncated by pre-defined rules to reduce the

variety of word types in word stemming, but these techniques add no information to models that improves the accuracy of analysis. However, computerization in content analysis is still very limited in media studies despite the increased availability of electronic data. Several factors explain the slow adoption of computerized content analysis for media research, including issues of copyright protection of the media sources and qualitative orientation of the field, but the greater content diversity of news articles explains a lot: content analysts face greater challenges in computerization, because news articles discuss a wider range of topics and views than political documents. The diverse topics and opinions in news stories result in a larger variety of vocabulary, thus most documents share only up to 1-5% of words with other documents (Phan, Nguyen, & Horiguchi, 2008). Add-one smoothing and word stemming have been commonly used to address the data sparseness in computerized content analysis (Manning, Raghavan, & Schütze, 2008). All the words are given the pseudo count of 1 regardless of the patterns of their occurrences in the data in add-one smoothing, or words endings are mechanically truncated by pre-defined rules to reduce the variety of word types in word stemming, but these techniques add no information to models that improves the accuracy of analysis.

Content analysists face the greatest challenge when they apply supervised-learning techniques (e.g. naive Bayes classifier) to media content, because they have to create very large training set to compensate for the data sparseness (Manning et al., 2008; Phan et al., 2008). While each document contains only a fraction the total vocabulary, whose distribution is highly concentrated to commonly used words that have little substantive meanings (e.g. function words), a training set must be comprised of thousands manually labelled documents. In recent years, online crowd sourcing made large scale manual content analysis possible, but it is not always the solution for content analysts, because it requires payment to the participants; even if

funds are available, it is difficult to recruit participants for non-English materials. Content analysists face the greatest challenge when they apply supervised-learning techniques (e.g. naive Bayes classifier) to media content, because they have to create very large training set to compensate for the data sparseness (Manning et al., 2008; Phan et al., 2008). While each document contains only a fraction the total vocabulary, whose distribution is highly concentrated to commonly used words that have little substantive meanings (e.g. function words), a training set must be comprised of thousands manually labelled documents. In recent years, online crowd sourcing made large scale manual content analysis possible, but it is not always the solution for content analysis, because it requires payment to the participants; even if funds are available, it is difficult to recruit participants for non-English materials.

For these reasons, it seems that the dictionary-based method, which is based on a long list of manually chosen words, has been most popular in computerized content analysis to date. For example, Roberts and McCombs (1994) identified media agenda in newspaper articles; Kellstedt (2000) analysed the framing of racial issues in *Newsweek* in terms of egalitarianism or individualism. More recently, Segev and Miesch (2011) identified the framing of 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. Many content analysis dictionaries are also made publicly available: the General Inquirer Dictionary (Stone, Dunphy, Smith, & Ogilvie, 1966), LIWC (Francis & Pennebaker, 1993), the Regressive Imagery Dictionary (Martindale, 1975) or DICTION (North, Lagerstrom, & Mitchell, 1984). Yet, adoption of these off-the-shelf dictionaries raises concerns regarding the validity of measurements, due to a lack of transparency in the dictionary making procedure (Grimmer & Stewart, 2013; Neuendorf, 2002).

This paper presents a new content analysis technique called Latent Semantic Scaling (LSS), developed by the author to perform large scale content analysis of English and Russian news articles for recently published studies (Author 2017a, 2017b, 2017c). LSS performs document scaling tasks in a similar way as Wordscore (Benoit & Laver, 2003), but it exploits the efficiency of vector space models (VSMs) to extract semantic information automatically from a large corpus (Turney & Pantel, 2010). Although most of the VSMs, such as latent semantic analysis (Landauer & Dutnais, 1997) and latent Dirichlet allocation (Blei, 2012), are unsupervised techniques, the LSS turns them into semi-supervised or fully-supervised machine learning techniques for theory-driven social scientific content analysis.

In the first half of this paper, I will demonstrate the possibility to apply VSMs to document scaling tasks by accurately measuring positive-negative tones of news stories on democracy or sovereignty in Ukraine with a semi-supervised technique (c.f. Author 2017a). To highlight its advantage over existing computerized content analysis techniques, I will also apply Lexicoder Sentiment Dictionary (Young & Soroka, 2012) to the same task. In the second half, I will explain how VSMs can be turned into fully-supervised document scaling model. To highlight its advantage over existing computerized content analysis techniques, I will also apply Lexicoder Sentiment Dictionary (Young & Soroka, 2012) to the same documents. In the second half, I will explain how VSMs can be turned into fully-supervised document scaling model by analysing Russian-language news stories in terms of framing of street protests (c.f. Author 2017b). The dimension for this document scaling is more complex than sentiment: whether street protests are framed as freedom of expression or social disorder. I will also apply Wordscore to the same task to show the strength of the VSMs over Bayesian models in analysing media content.

VSM for document scaling

VSMs are very useful in analysing media content, because they automatically extract semantic information from a corpus (Turney & Pantel, 2010), but their raw outputs are no more than vector representation of words. Therefore, there are two steps in LSS before applying a VSM to document scaling: (1) selecting features relevant to the subject of interest, and (2) specifying an axis of the document scaling. The feature selection is performed with *target words*, which express concepts that the researcher focuses on, and the specification of the axis is achieved with *seed words*, which define dimensions that the researcher wishes to measure. Here, words frequently used in conjunction with target words are selected as features, and the feature words used in similar contexts as seed words are given large polarity scores. This is a combination of syntagmatic and paradigmatic analyses, which have been treated as two different approaches to automated synonyms extraction in literature (Schütze & Pedersen, 1993; c.f. Turney, 2001; 2003). This combination is advantageous, because syntagmatic and paradigmatic analyses capture very different types of semantic relations of words.

In actual implementation of LSS, syntagmatic analysis is performed as collocation analysis with fixed word windows, and paradigmatic analysis is as latent semantic analysis (LSA) with the cosine similarity measure. The document scaling model is essentially a large set of subject-specific feature words with continuous scores representing their polarities. Although the structure is very simple, it is dissimilar to the products of lexicon expansion techniques, which have been developed by computer scientists to generate a large lexicon automatically from a small set of pre-defined words (c.f. Liu & Hu, 2004). With the continuous scores given to feature words, LSS can adopt Wordscore's document scoring method, making the results more accurate than that of dictionary-based content analysis.

The most straightforward application of LSS is sentiment analysis of news on specific subjects, because a set of general positive-negative seed words is already available (Turney & Littman, 2003). Given the seed words, only manual input to LSS is target words that is usually a wildcard expression of target concept. I will explain the detail on how the document scaling technique works through construction of modelsa model that measures positive-negative tones in news stories on Ukraine's democracy or sovereignty. This models areThis model is utilized in longitudinal analysis of English news stories published by Russian news agency, ITAR-TASS, during the 2014 Ukraine crisis (Author 2017a, 2017b).

For the construction of the model, I created 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. As pre-processing, I segmented all the news articles into sentences, and removed all the capitalized words. Sentence segmentation is necessary to prevent collocation analysis and LSA from being affected by preceding and succeeding sentences; removal of capitalized words is required to limit the impact of proper names and adjectives on LSA in inferring general meanings of words.

Feature selection

LSS utilizes collocation analysis to select features that are strongly associated with democracy or sovereignty. The system identifies words frequently co-occurred near target words in the corpus. The target words 'democra*' are 'sovereign*' and windows size is set to 10. The level of association with target words are measured by the likelihood ratio statistic, or g-score (Hoey, 2012). To compute g-scores, the system counts 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 constructs contingency tables:

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

With these tables, the system calculates g-scores g_i for w_i by comparing observed counts n_j with expected counts e_j, which are estimated by the marginal distribution of the observed counts in the same way as a chi-square test:

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

The system selects up to 1,000 features with $g_i > 10.83$, the critical value for a 99.9% confidence level if their observed counts are greater than their expected counts, $n_1 > e_1$. The number of features selected by these criteria is 778 for democracy and 626 for sovereignty. All the features for democracy seem to be intuitively related to the topic, but there are a few financial words such as 'debt' and 'bonds' also present (Table 1). These financial words are erroneously selected, but they have a limited impact on the final outcomes of the 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

Word scoring

The system estimates sentiment of the feature words based on their distances to seed words in a latent semantic space. The seed words are a set of general English words that express positive or negative sentiment {good, nice, excellent, positive, fortunate, correct, superior} and {bad, nasty, poor, negative, unfortunate, wrong, inferior} (Turney & Littman, 2003). Although the semantic proximity between words can also be calculated in a raw term-document matrix, a latent semantic space enables much more accurate estimation, because singular value decomposition (SVD) of a term-document matrix reduces noises and sparseness, only leaving essential semantic information (Turney & Pantel, 2010). Although earlier studies suggest that feature weighting by tf-idf or PMI improves the accuracy of similarity computations in VSMs, feature weighting is not performed in LSS as it does not improve the result.

The term-sentence matrix X, which is derived from a large corpus and consisting of 270,000 rows and over a million columns, is too noisy and too sparse to estimate semantic proximity (left in Figure 1). SVD decomposed the matrix into three matrices, U, D and V, and constructed a matrix $\hat{S}$ with only 300 columns (right in Figure 1):

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

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

With the matrix $\hat{S}$, the system estimates the sentiment of words taking the mean of cosine similarity to each of the seed words: the sentiment score v_i for a word w_i is a mean cosine similarity to seed words weighted by seed scores p_j, which are simply +1 for the positive seed words and −1 for the negative seed 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 $\hat{S}$.

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

Figure 1: Notional illustration of dimension reduction by SVD

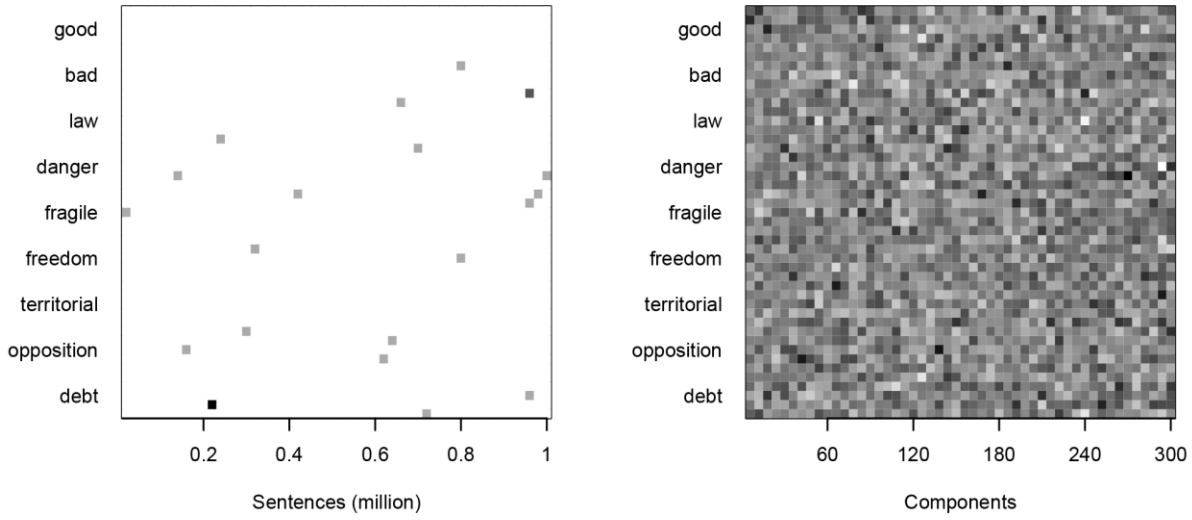


Figure 2 illustrates this word scoring method in a semantic space, in which the positive-negative dimension (the dotted line) is defined by two sets of seed words, and word w_i and w_2 are located in different distances from positive seed words but in the same distance from negative seed words. Due to the greater proximity to the positive seed words, w_1 gains higher score than w_2, reflecting its importance as a word precisely on the positive-negative dimension.

Figure 2: Concept of spatial word scoring in LSS

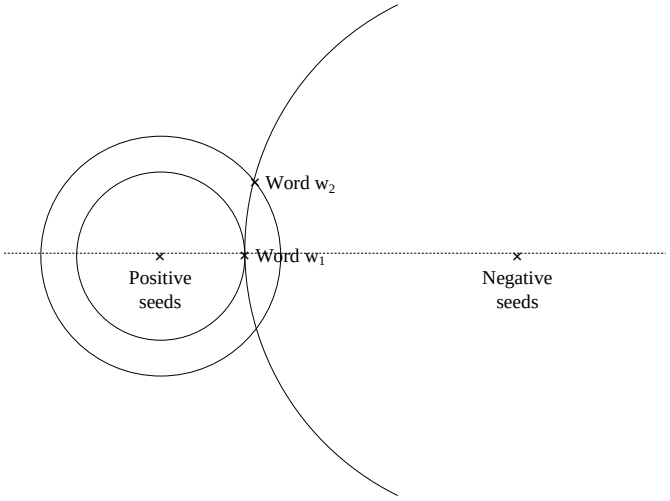


Table 2 and 3 present the top 20 most positive and negative words on democracy and sovereignty. In both tables, many of the words are intuitively positive or negative, but some are not. For example, ‘intensify’ in democracy, is not always used positively, but we cannot judge if its score is accurate, unless we investigate its usage in the large corpus. Also, ‘upon’ in sovereignty is a function word lacking substantive meaning, but it could be removed easily if a larger list of stopwords were to be utilized. More importantly, roughly the same scores given to different forms of the same words (‘strengthening/strengthen’ and ‘strong/strongly’) indicate that the estimation of the sentiment the words is accurate (Manning et al., 2008).

Table 2: Most positive and negative words for democracy

Rank	Positive	Score	Negative	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 words for sovereignty

Rank	Positive	Score	Negative	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 are given to the feature words, the LSS model become ready for content analysing news articles. When words $w_{i...I}$ occur in a story in total of m times, v_i is their sentiment score, and f_i is the frequency count of the words w_i, sentiment of news articles, or the document score, d' is computed as:

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

Validation

I applied the document scaling model to two samples of news stories (democracy or sovereignty) on Ukraine published by the Russian agencies (ITAR-TASS and Interfax) to test their validity. I also applied a Lexicoder Sentiment Dictionary, which was created originally to

analyse Canadian newspapers' coverage of elections, to the samples as an example of an off-the-shelf dictionary.[1] In dictionary-based content analysis tools, including Lexicoder, the sentiment score of a document d is the difference between the 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 3 and 4 compare document scores assigned by computerized content analysis (Lexicoder or LSS) with those assigned by manual content analysis.[2] In stories on democracy, many scores computed by Lexicoder are accurate, although the overall correlation is only moderate (r=0.46) due to the overestimations of positivity (#6, #16) or negativity (#26). In LSS, there are two large errors (#8, #27), but other documents are accurately scored, achieving stronger correlations with human scores (r=0.77). In sovereignty, however, Lexicoder underestimates the positivity of many documents, but extreme cases (#6, #21, #25) are scored very accurately, hugely affecting the correlation coefficient (r=0.65). LSS is less accurate in sovereignty than in democracy, creating random errors around the regression line, but it still outperforms Lexicoder in this subject (r=0.70).

Figure 3: Sentiment of news on democracy

[1] Lexicoder has a negation words handling mechanism, but I adopted the common bag-of-words approach.
[2] Raw scores of LSS are rescaled to between −100 and +100. In manual coding, I classified on a five-point scale {1: very negative, 2: negative, 3: neutral, 4: positive, 5: very positive}, and calculated document scores by taking means of sentence scores.

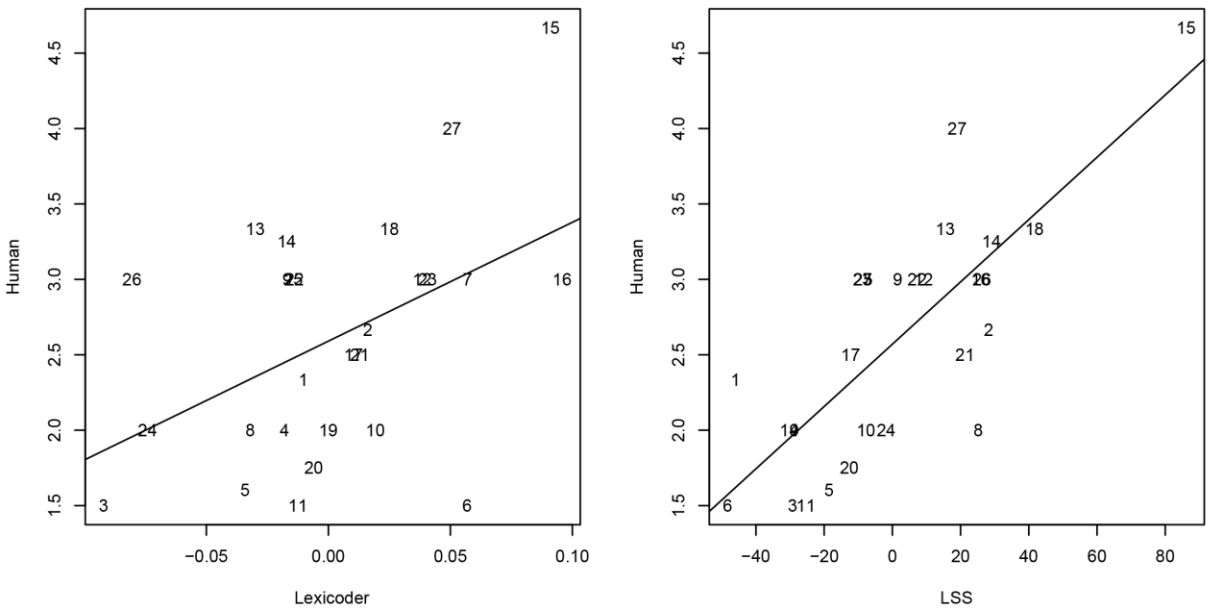
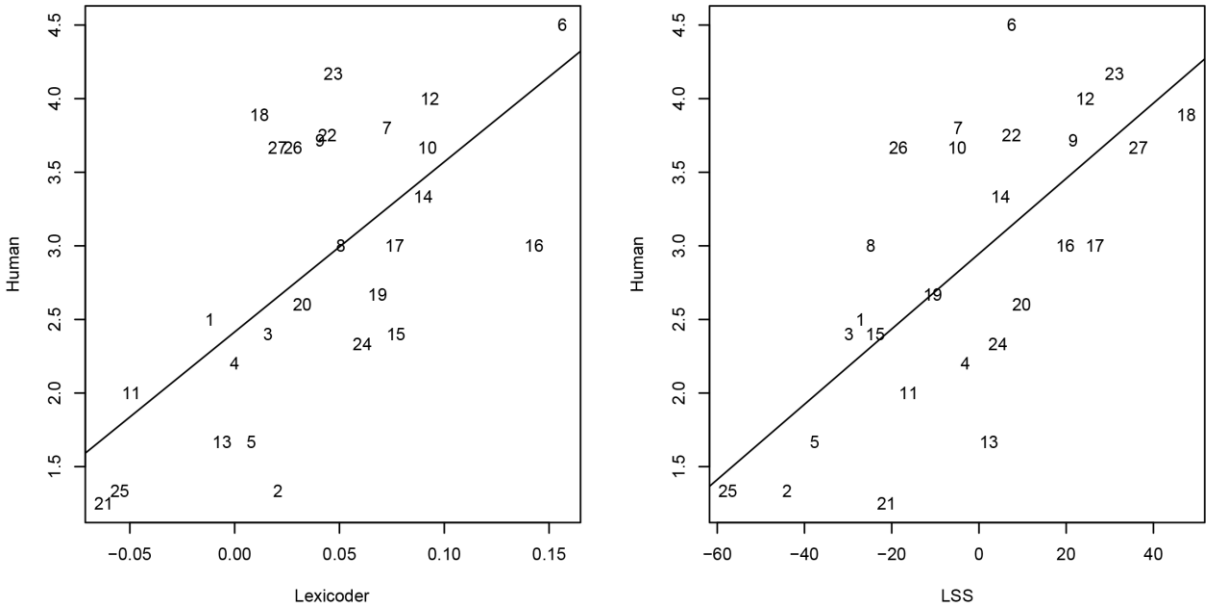


Figure 4: Sentiment of news on sovereignty



Supervised-VSM document scaling

I have constructed document scaling models on sentiment with existing seed words, but nothing prevents researchers from selecting seed words themselves and applying them for their own purposes. However, manual selection of seed words is sometimes difficult to achieve, particularly when one wishes to measure complex dimensions in news content. An example of

complex dimensions is media's framing of street protests as 'freedom of expression' vs. 'public disorder' (Author 2017c). Therefore, I will present a fully-supervised VSM document technique which is based on an algorithm that select seed words automatically guided by manually coded documents.

In the supervised-LSS, a larger corpus is supplied as external data to aids estimation of parameters for words that do not or only rarely appear in a training set. As a similar approach, Nigam et al (2000) proposed an iterative algorithm based on naive Bayes classifier, in which a supervised model is trained on both machine-classified and human-classified documents to improve the reliability of parameter estimation. Phan et al (2008) incorporated parameters estimated in Wikipedia pages by LDA models into a maximum entropy model trained on manually classified documents for the same purpose. Although the techniques proposed by the computer scientists are very complex, fitting VSMs to training data in LSS is achieved by a simple forward stepwise algorithm, maximizing correlations between manually scores and estimated scores given to documents in training set.

The external data is a corpus of news articles constructed from the same source as the training and test sets. I constructed a corpus of Russian-language 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 (Author 2017c). The corpus contains 39,787 full-text news articles or transcripts of news casts on street protests. As before, I segmented news articles into sentences and eliminated all proper nouns and adjectives from the corpus.

I 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 framing the

protests ‘explicitly as social disorder’ to ‘explicitly as freedom of expression’. I aggregated the sentences scores to obtain accurate document scores, and then allocated the first half of the documents for a training set and the last half for a test set.

Fitting VSM model

The goal in fitting a VSM to training set is discovering 5 to 10 pairs of polarity words that define the freedom-disorder dimension. Since there are too many types of words in the large corpus, two criteria are applied initially to select candidates. First, candidates are only those strongly associated with the target words, which are identified by the collocation analysis. This is in fact the same criterial as for the selection of feature words. I selected candidates from the top 10% of 10,380 types of words that occur at least 5 times within a 10-word window from “protest” (“протест*”) in the Russian news corpus.

Second, candidates are only those strongly correlated with the dimension without being paired with other candidates. To test the level of correlation with the freedom-disorder dimension quickly, the system calculates pair-wise cosine similarities between all the top features in a SVD-reduced matrix $\hat{S}$. When cosine similarities for all pairs are stored in a symmetric matrix D, it has $K = 1{,}038$ columns and rows corresponding to the seed candidates $c_{k\ldots K}$. Within the matrix D, scores for words are found in a kth row or column vector when word c_k is the candidate.

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

The system weights them by normalized frequency of words in documents in the training set to obtain temporary document scores (Equation 1.5), and records their correlation with manual scores as r_k. These correlation coefficients allow the system to prioritize the candidates in the stepwise selection process. I selected only 50 seed candidates with the largest absolute

correlation coefficient from both sides of polarity in the study. Their seed scores p_k are assigned in the following manner:

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

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

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

This adjustment is not only to equalize document scores obtained from the varying similarity vectors d_k, but also to increase the dispersion of seed words in the semantic space. A candidate's high average similarity to other candidates indicates that it is in a dense cloud of candidates in the semantic space, but ideal seed words are separated in the space, covering wide regions of the semantic space. Seed words widely dispersed across the semantic space less likely to overfit to the training data.

With the top seed candidates from both polarities, the system constructs pairs of seed words $\{c_k, c_l\}$, identifying a partner c_l for c_k such that (1) the partner has an opposite polarity, $p_l \neq p_k$; (2) the model $d_{\{k,l\}}$ yields a higher correlation coefficient than before the paring, $r_{\{k,l\}} > r_k$ 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 largest absolute correlation coefficient $|r_k|$, all other seed candidates enters this stepwise paring process. This process continued until at least five pairs had been found, and new pairs started decreasing the overall correlation (this takes around 30 seconds on a laptop computer).

Table 4 shows seed words that are automatically selected through the stepwise selection process. Although it is difficult to judge the suitability of seed words without being familiar with the contexts in which they are used, the freedom-of-assembly seeds seem to be related to legal or administrative procedures, while the social-disorder seeds seem to describe either the attributes or behaviour of protesters.

Table 4: Automatically selected freedom-disorder 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	fundamental	-48.9
использования	utilization	-98.1
подчиняющиеся	obeying	-130.2
госпереворот	coup	-149.0
нежелания	unwillingness	-306.0

Validation

I applied the LSS model and Wordscore to the 15 manually coded news articles in the test set for comparison. In Wordscore, when there are H manually scored documents, the score for a word v_i is its average frequency weighted by document scores d_h in training set:

$$v_i = \sum_{h}^{H} \frac{f_i}{k_h} \cdot d_h \qquad (2.4)$$

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

Figure 5 shows document scores assigned to the training set by LSS and Wordscore. Wordscore reproduces scores assigned by human coders (r=0.93) much better than LSS (r=0.85); 95% confidence intervals are also very small in Wordscore, indicating high confidence in

estimated scores. However, LSS (r=0.76) performs much better than Wordscore (r=0.39) in the test set (Figure 6). While LSS's confidence intervals are nearly the same size as in training set, they became much larger in test set in Wordscore. This suggest that the Wordscore is affected by sparseness of the data, only recognizing only few features in the test set.

Figure 5: Freedom-disorder framing of protests in Russian news (training set)

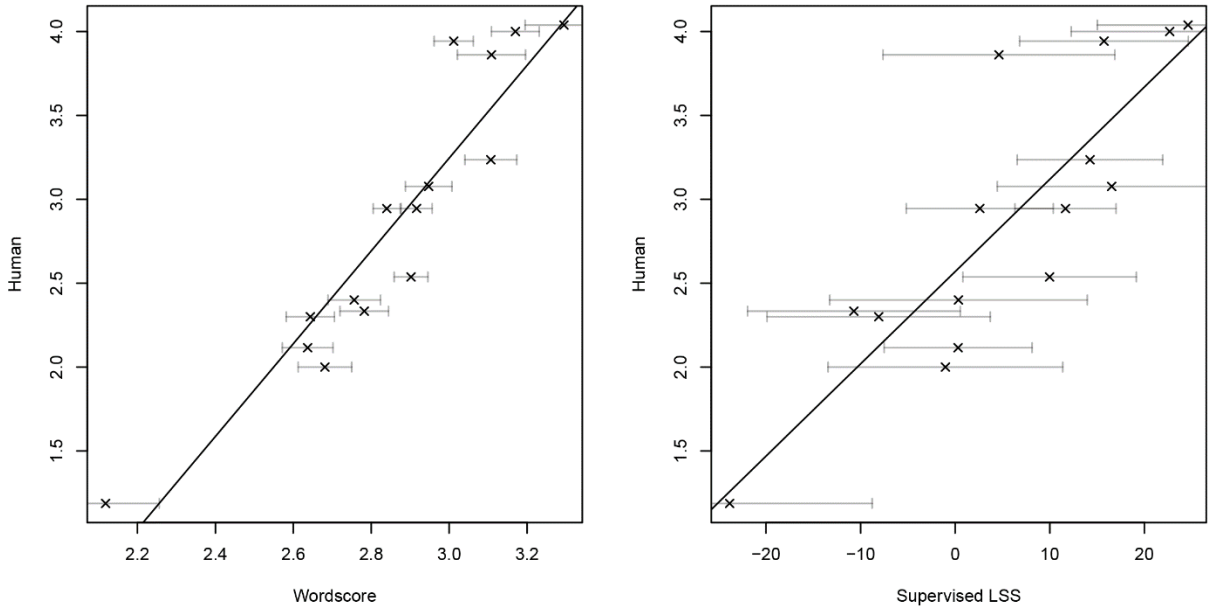
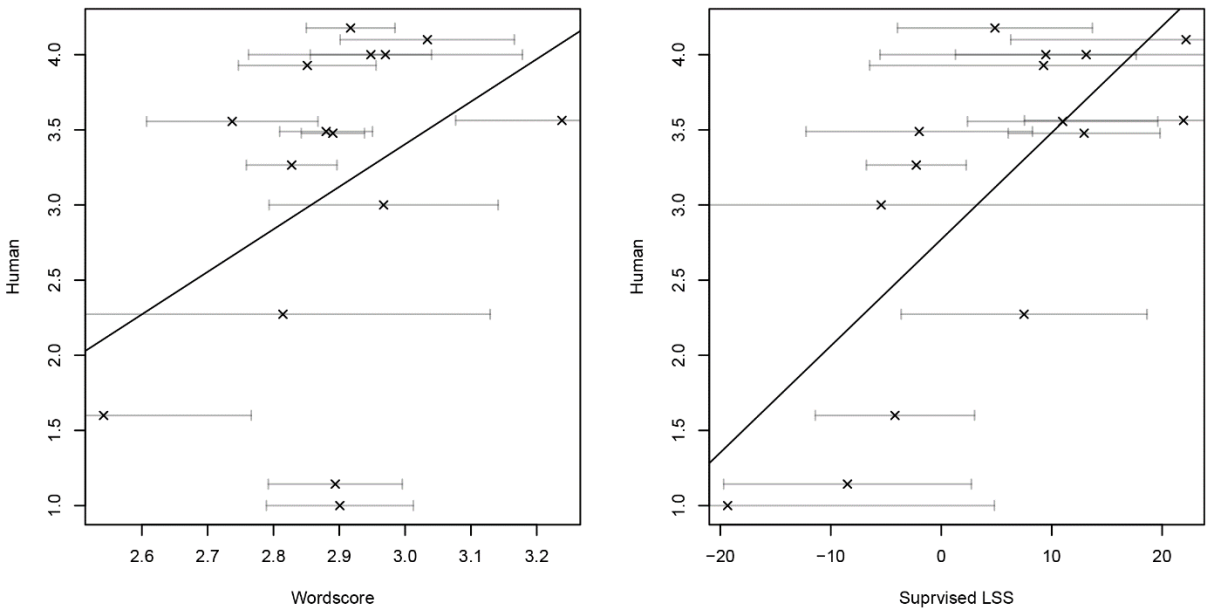


Figure 6: Freedom-disorder framing of protests in Russian news (test set)



Discussion

In the first half of this paper, I demonstrated that VSMs can be easily applied to document scaling tasks as a semi-supervised machine learning technique: the only additional manual input was target words in sentiment analysis. The result of the document scaling was as good as or even better than the manually compiled sentiment dictionary. The LSS model performed better in news stories on democracy in Ukraine, because the system extracted words used by the Russian news agencies from the corpus, which are different from those used by the North American news media. This means that content analysts can construct subject-specific document scaling models for sentiment analysis without additional costs in every research projects, avoiding use of off-the-shelf dictionaries that raises validity concerns when applied to new subjects (Grimmer & Stewart, 2013).

Nonetheless, there was a weakness of LSS models vis-à-vis the manually compiled dictionary: the models were less accurate in scoring extremely positive or negative news stories than the dictionary. This is due to the LSS's scoring method, where polarity of a words is calculated based on its proximity to seed words representing the both ends of the scale, and

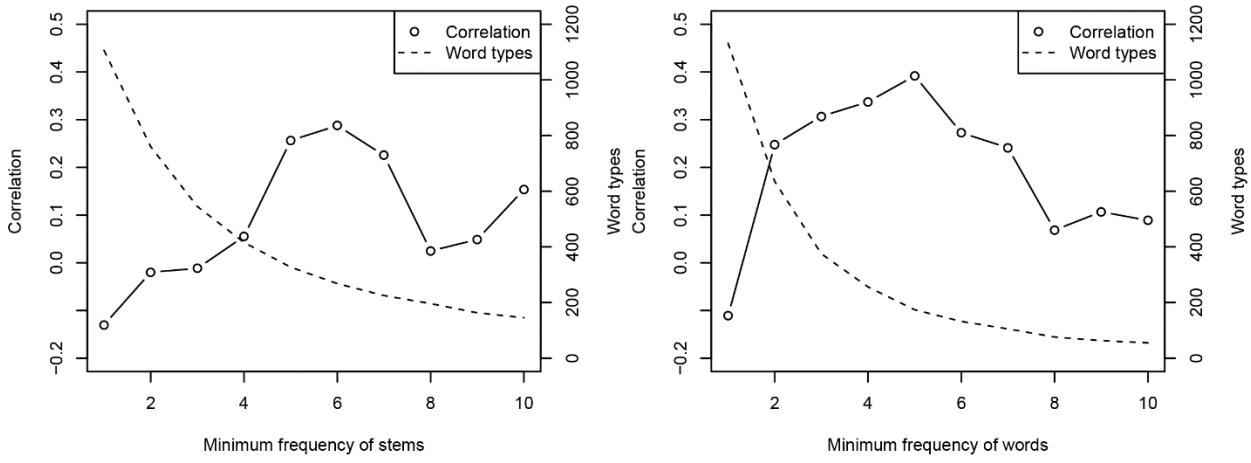
extreme words beyond the range do not gain large enough polarity scores. Although the English sentiment seed words are widely applicable, their positivity or negativity is still moderate compare to the very strong words appeared in Russian newswires. This encourages us discover extra sentiment words to make sentiment analysis more accessible.

In the second half of the paper, I have shown that VSMs can also be employed in a fully supervised machine learning technique. Despite the lexical diversity in the Russian news stories, LSS replicated manual coding well only with 15 document in the training set. This efficiency is owing to the VSM's ability to extract semantic information from large corpora fully automatically: all features are selected based on syntagmatic association with the target words, and they are scored based on paradigmatic proximity to each other in the larger corpus. Fitting the VSM to training data is achieve by automatically selecting seed words for the given dimension with the novel stepwise algorithm. Unlike the earlier techniques (Nigam et al., 2000; Phan et al., 2008) that utilizes external data to expand models, no parameter is directly estimated from the training set and thus there is no need of complex mathematical operations to combine parameters from training and external data with the algorithm.

The comparison between LSS and Wordscore has shown the challenge that diverse content of news articles poses for researchers: a very large number of documents must be manually coded for supervised techniques that estimate parameters directly from a training set. In my attempt to construct the best-performing Wordscore model ($r=0.39$), I had to eliminate words that do not occur more than five times to improve the reliability of parameter estimation. As shown in the right panel of Figure 6, when all the words in the training set are included in the model, the model does not replicate human scoring at all ($r=-0.11$); when words occurring only once in the training set are excluded, the correlation increased 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 7).

Figure 7: Minimum frequency of words and types of words



Nonetheless, introducing the higher threshold for the minimum frequency rapidly decreases the number of word types found in both the training set and test set (the broken lines in Figure 6): there are 1,132 types of words when there is no threshold, but that number halves when the minimum frequency is set to two. When the threshold is raised to five, only 175 types are left. After this point, the correlation starts falling sharply as the model fails to recognize relevant features in the test set. This is a bias-variance trade off in a Bayesian model, and the only solution for the model is to increase the size of training data. LSS can also be affected by this type of trade off, but it is easily solved by providing a large corpus without expensive manual coding.

Stemming is a commonly-used technique to compensate for data sparseness, but poorer performance of Wordscore when stemming is applied indicates that it is not a solution, being potentially harmful (left in Figure 7). The better practice is estimating parameters separately for

different forms, and assign similar values to those have the same meanings (Manning et al., 2008). This is exactly what LSS does by estimating semantic relations between words in large external data with VSMs. Their ability to estimate semantic relations accurately is clearly shown in the very close scores given to “strengthening/strengthen” and “strong/strongly”.

Overall, the successful application of the VSM to content analysis is owing to the introduction of target words in addition to seed words. In the earlier study, collocations analysis and LSA are parallel approaches to synonym extraction (Turney, 2001; Turney & Littman, 2003), but this study made clear that these are suitable for different purposes. Collocation analysis effective in extract subject-specific terms, whereas LSA is accurate in identifying synonyms. Both methods are highly scalable, easily extracting and scoring subject-specific features with high confidence in corpora that are comprised of tens of thousands of news stories.

Finally, only LSA is utilized as an undying model for LSS in this study for its accessibility, but nothing prevents us from adopting other VSMs (Turney & Pantel, 2010). Since the emergence of LSA in 1990, other models such as NMF (Lee & Seung, 2001) and LDA (Blei, 2012) models have appeared; more recently Mikolov et al. (2013) have propose an efficient vector representation of words known as Word2vec. These VSMs are based on different distributional assumptions and constraints, but can potentially be used as an underlying model for LSS to estimate semantic relations of words more accurately. Therefore, I encourage readers to apply those models to improve the accuracy of computerized content analysis of media content.

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. https://doi.org/10.1080/07907180312331293249

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

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. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 1–31. https://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. https://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

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562).

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.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [Cs]*. Retrieved from http://arxiv.org/abs/1301.3781

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

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents Using EM. *Mach. Learn.*, *39*(2–3), 103–134. https://doi.org/10.1023/A:1007692713085

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

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 91–100). New York, NY, USA: ACM. https://doi.org/10.1145/1367497.1367510

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

Schütze, H., & Pedersen, J. (1993). *A Vector Model for Syntagmatic and Paradigmatic Relatedness*.

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. https://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.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Retrieved from http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=5765594

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. https://doi.org/10.1145/944012.944013

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188.

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