Rethinking sentiment analysis and 'master narratives': an alternative unsupervised text analytics approach using 'information space differences'

Peter A. Chew, Jessica G. Turnley
Galisteo Consulting Group, Inc.
4004 Carlisle Boulevard #H
Albuquerque, NM 87107, USA
{pachew,jgturnley}@galisteoconsulting.com

Abstract

Widespread interest exists in government in applying data analytics techniques to 'big data', for example social media data, to gauge sentiment among populations. Two catchphrases which have ridden this wave of interest and gained currency are 'sentiment analysis' and 'master narratives'. However, both entail methodological problems, and we suggest a rather different way of looking at big data, specifically, social media data, which can be computationally well-defined and validated, avoiding reliance on heuristics, and which is fully driven by the data itself rather than any analyst preconceptions. We show via examples that our unsupervised alternative is in fact capable of teasing out themes related to sentiment and master narratives from big social media data – themes which could potentially be useful in informing diplomacy.

1 Introduction

With the globalization of the internet, interest in organizing and making sense of content is growing constantly. One area studied for around a decade and a half is **sentiment analysis**, e.g. (Pang et al., 2002). Sentiment analysis aims to automate answering the question of whether people express positive or negative sentiment, usually with respect to some topic, in text. Sentiment analysis promises to provide a low-cost alternative to opinion polling; people voluntarily express sentiment online, and all that data is simply there for the taking for anyone who can mine and make sense of it. For governments concerned with information operations among foreign populations, the ability to mine those populations' online posts is very attractive when compared to the alternative – the traditional, manual approach of conducting surveys, which usually requires a presence on the ground.

Another framework which could contribute to sensemaking in the context of sentiment and big data is the **master narratives** framework, e.g. (Halverson et al., 2011). According to Halverson et al., 'a master narrative is *a transhistorical narrative that is deeply embedded in a particular culture*' (p. 14). As an example, they cite the master narrative of 'Pharaoh'. Pharaoh was an Egyptian tyrant mentioned in the Bible and Quran, but more recently has been used as a moniker for former Egyptian President Hosni Mubarak. The narrative is 'deeply embedded in Islamic culture' because of its connection to the Quran, the most sacred text of Islam. It is 'transhistorical' both because of the Quran connection and its *application* in a modern context. By applying the title of 'Pharaoh' to Mubarak, bloggers were able to evoke the constellation of evil behaviors surrounding the Quranic figure –thus expressing their sentiment about Mubarak in shorthand.

With this brief background, the stage is set for stating our own goals: to develop a robust computational framework to make sense of social media data, as sentiment analysis and master narratives do, but avoiding the problems (discussed below) in those approaches.

Under the general heading of 'sense-making', our approach extracts the principal topics inherent in the data, the subject of prior work (Chew, 2015). It thus allows us to automatically identify ways in which sets of topics extracted from a corpus differ from

each other. As we will see later, our approach is able to show differences with significance in English language and Russian language tweets about the same real world referent (NATO exercises, in our example). We believe that divergences between the English and Russian 'information spaces' (different ways of making sense of the referent) revealed through this multilingual topic analysis can be very useful to those interested in information warfare or analyses of geopolitical players and events.

To make things more challenging, but also bringing assumptions closer to real-life, we assume the social media data in question is multilingual, and that we want our model to work equally well regardless of topic or language mix. It is a reasonable assumption that each dataset with which an analyst works will differ with regard to topic, how sentiment is expressed, languages in which people post, and other dimensions.

And if all that were not challenging enough, we insist that any model we develop should be capable of empirical validation.

The approach we propose, unlike most sentiment analysis, is completely **unsupervised.** Patterns emerge from (rather than being imposed from extraneous experience on) data. This is similar in approach to ethnomethodology, an approach to social data analysis which is theoretically agnostic (Garfinkel, 1967). It also grows organically out of our previous multilingual topic-extraction work (Chew, 2015) based upon vector-space text analytics. Although our approach produces results of a kind different to those of sentiment analysis, and thus the two cannot be directly compared experimentally, we demonstrate that our approach does provide output usable for inferring useful and potentially actionable conclusions about sentiment with respect to particular topics.

The paper is organized as follows. Section 2 outlines some of the weaknesses and difficulties with sentiment and 'master narratives' analysis. Section 3 outlines our alternative approach. In section 4, we describe how our approach can be empirically validated and present relevant results. In Section 5, we describe an application of our approach to around 200,000 Twitter posts in Russian and English relating to May 2016 NATO exercises in Eastern Europe; we show how we were quickly able to learn useful and non-obvious facts about differences between Russians' and Westerners' views of the same events. Finally, section 6 concludes on our findings.

2 Sentiment analysis and master narratives: shortcomings and difficulties

2.1 Sentiment analysis

Sentiment analysis is usually approached in one of three ways: either (1) 'top-down' with supervised machine learning, e.g. Pang et al. (2002), (2) 'bottom-up' using lists of 'sentiment-bearing' words, e.g. (Kim & Hovey, 2004), or (3) some combination of the two, e.g. the semi-supervised approach of (Sindhwani & Melville, 2008).

Approach (1) involves manually labeling examples of text, then using machine learning to infer sentiment of previously unseen text. For example, in Pang et al., movie reviews are given star ratings according to a Likert scale by users who also write textual reviews. Each star rating summarizes how (un)favorably the user rated a given movie. Pang et al. then use a standard vector-space model in which the words of the text are the features of each review, and apply a support-vector machine (SVM) to extrapolate favorability for unseen reviews, with up to 81.6% accuracy.

Under approach (2), individual words are labeled with respect to sentiment. Again, the words in the text can be treated as features of each chunk of text, and a machine learning algorithm can be set up to infer the sentiment of the text as a whole from the sentiment of all its constituent words.

Each approach has its drawbacks. Supervised learning techniques require tagged training data (Vapnik, 1998) representative of the test data. Thus, it is hard to see how a sentiment classification algorithm trained, say, on movie reviews would necessarily be applicable to the opinions of a population with respect to NATO exercises.

The shortcomings of the bottom-up 'sentiment-bearing words' approach was commented on by (Beasley & Mason, 2015), who find low correlation between a sentiment 'dictionary' (Linguistic Inquiry Word Count) and a 'well-validated scale of trait emotionality'. Perhaps because of the fluidity of language, it is hard objectively to pin down the precise 'sentiment' of individual words, or even to define sentiment in the first place.

Further, both types of approach to sentiment analysis are limited with regard to multilingual text. The problem of non-generalizability of supervised approaches is magnified when it comes to multilingual content. Algorithms trained on, say, English will not

be applicable to Russian, and appropriate multilingual training datasets will be even less available than monolingual ones. With 'sentiment-bearing words' approaches, the difficulty of pinning down sentiment by word is magnified. Either one must compile one word-list per language – which requires language experts – or one must assign sentiment to words and then translate (CASOS, 2016), in which case any uncertainties about the correct sentiment for a word are amplified by translation. The assumption that sentiment can be assigned to words cross-lingually in this way seems highly simplistic.

2.2 Master narratives

At the outset, we should state we are not aware of attempts to marry the master-narratives and data-analytics frameworks. This may be in part because the academic literature on master narrative theory (e.g. Halverson et al., 2011) is sparse, and relatively recent. At this writing, Google Scholar shows only about 25,000 hits for 'master narratives', with almost half appearing since 2010. The national security literature database, dtic.mil, shows only 20 hits.

Another reason master narratives analysis may not have been applied to big data is simply that those who have applied the construct have traditionally worked a different way, heavily relying on subject-matter experts and compiling lengthy prose reports. This entails significant costs in both time and money: examples are both (Halverson et al., 2011) and 'narrative analytics' work carried out by Monitor 360[1].

As we began to think about how to implement master-narratives analysis computationally, we realized several things were lacking:

- A robust definition of 'master narrative';
- A robust statement of the problem the computer was supposed to solve (identifying master narratives for the analyst? identifying where known master narratives are referenced?)
- And even if we could formulate a clear problem statement, how would success in solving the problem be recognizable?

Although future work may answer these questions, for now at least, we believe it is an open question how big data and computational modeling techniques can directly access and utilize master narratives.

3 Our approach: look at differences between 'information spaces'

3.1 Data

The dataset we worked with for this paper is a collection of 206,734 Twitter posts, gathered from the Twitter 'garden hose' by specifying 'NATO' and the Russian equivalent 'НАТО' as search terms. All posts are from between May 5, 2016 and June 7, 2016, with English and Russian posts distributed approximately evenly across the time period. Statistics for this corpus are as follows:

Language	# docs*	Types	Tokens
EN	127,220	170,347	2,116,052
RU	79,514	105,589	1,113,833
Total	**206,734**	**275,936**	**3,229,885**

Table 1. Statistics by corpus and language

3.2 Overview: methodological novelty

Instead of trying to answer, say, the question 'what is the sentiment of Russians about NATO?' (the question sentiment analysis poses), or 'what master narratives do the Russians use about NATO?', we seek simply to answer the following questions, which we see more fundamental, neutral and framework-independent than those above:

- What themes are discussed in all 206,734 posts?
- What themes are weighted towards either Russian or English?

The answer to the second of these questions tells us, simply, how the Russian-language discourse *differs* from the English-language discourse and vice-versa, characterizing what we have called two different **information spaces**.

3.3 Use of standard LSA, with a key change

Expressed in these terms, our problem is one that already has a solution with a good pedigree. That solution is Latent Semantic Analysis (LSA) (Landauer & Littman, 1990), which can be used to extract topics, each corresponding to a principal component in Singular Value Decomposition [SVD], the algorithm underlying LSA. The only difficulty,

[1] http://www.monitor-360.com/narrative-analytics.

which is not trivial, is in adapting LSA to take *multilingual* text as input and produce as output a list of topics weighted by language – such that the *same* set of topics spans the different languages. By this we mean, for example, that if one of the topics that emerge from the 206,734 Twitter posts is 'NATO missile shield', then posts in both Russian and English should score highly on that topic. We look for topics which are the focus of attention for both Russian and English posts – and then for the topics 'left over.' It is the leftovers that offer clues to how the Russian and English discourse differ.

In our LSA framework, 'documents' are Twitter posts, and 'terms' are words separated by white space within the text. As is usual in LSA, token-frequency statistics are gathered into a term-by-document matrix X, each cell of which shows how many times a given term occurs in a given document. An entry *x* in row *i*, column *j*, for example, can indicate that term *i* occurs *x* times in document *j*. Also as customary in LSA, e.g. (Dumais, 1991), we weight X to discount the frequencies of non-distinctive terms; the weighting scheme we use is Positive Pointwise Mutual Information (PPMI). Unlike in 'standard' LSA, we perform an additional step to transform X into a 'multilingualized' version of X; see below.

The weighted, multilingualized matrix X is then factorized using SVD into three further matrices, U, S, and V, with truncation applied to the SVD: we discard all but the top 90 principal components, as follows:

$$\mathbf{X} \approx \mathbf{U}\mathbf{S}\mathbf{V}^T \qquad (1)$$

In mathematical terms, U is an orthonormal matrix of left singular term vectors, S is a diagonal matrix of singular values, and V is an orthonormal matrix of right singular document vectors (Golub & van Loan, 1996).

In this framework, U and V numerically encode the weighting of each term and document, respectively, in each 'topic'. U and V can also be thought of as a collection of term and document vectors in a single topic space. The purpose of our 'multilingualizing' step mentioned above is to ensure that U and V relate all documents, in whatever language, to the *same* set of cross-language topics. This means that Russian and English documents about the 'NATO missile shield', for example, will tend to be similar in the orthogonal space and score highly on similar concepts.

Finally, to measure the extent to which topics are weighted towards one language or another, we can partition V by language (since we know the language of each source document) and then calculate the sum of squares per topic, per language, giving a measure of 'amplitude' of each topic in each language.

3.4 Our innovation to 'multilingualize' LSA

Under standard LSA, if there is a mixture of Russian and English in the corpus, some rows in X will correspond to Russian and others to English terms. Likewise, some columns of X will correspond to Russian documents , others to English documents. Since it is likely that most documents are in just one language – i.e., Russian terms tend not to occur in English documents and vice-versa – X will look as in Figure 1, where non-zero entries are concentrated in certain quadrants (blocks) of the overall term-by-document matrix.

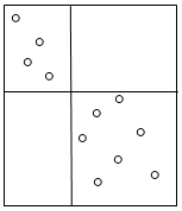

Figure 1. A corpus of Russian and English documents

The basic problem we must overcome is that with standard LSA, topics will be language-specific. To overcome this problem, we propose a mathematical/linguistic pre-processing step. In addition to the non-parallel corpus from which X is formed, we introduce a parallel (or multi-parallel) corpus. In the latter, all the languages that we anticipate encountering in X must be represented. Using a technique well-known in MT (Brown et al., 1994), we perform word-level alignment in the parallel corpus and then compute a set of translation probabilities for each source word, conditional upon source and target language. For example, we might compute that Russian 'нет' translates into English as 'not' with probability 0.6 and as 'no' with probability 0.4. These probabilities are gathered into a separate matrix which we call Y.

Since Y encodes the probability of any in-vocabulary term mapping to any other in-vocabulary term, Y will be a square matrix of size $n \times n$, where n is the multilingual vocabulary size. For out-of-vocabulary words (e.g., those in Twitter but not the parallel corpus), additional all-zero rows and columns are

added into Y. Probabilities are also included in Y to map each language to itself; in this case, the values in Y are 1 along the diagonal and 0 elsewhere, reflecting the fact that each term maps to itself in the same language with probability 1. If we are interested in analyzing documents in one language only, then Y will be the identity matrix and subsequent steps (described in the next section) reduce to standard LSA.

3.4.1 Computation of matrix product

Prior to SVD, we now compute the matrix product (YX), with weighting (e.g. PPMI) applied to X *prior* to computing the product. From a linguistic/mathematical point of view, this achieves 'multilingualization' of each document. Continuing our example above, any Russian document that included Russian 'нет' with weight 5 now *also* contains 'not' and 'no' with weights 3 and 2 respectively. This step transforms the matrix in Figure 1 so that the empty quadrants are filled in approximately as densely as the non-empty quadrants. Also prior to SVD, we transform (YX) by L2 normalization (Golub & van Loan, 1996), by document and 'translated' language. Mathematically, this ensures that each translation of each language has a vector in (YX) of unit length, and thus that each document and translation is in some sense weighted equally.

Note our approach here opens the door for the same word in one language to be translated multiple ways (with appropriate probabilities included), and for this all to be reflected in the linear algebra. This in turn allows more 'paths' for SVD to make appropriate semantic connections between languages; we view this as an advantage.

4 Empirical validation

Recall that our approach to finding differences between, say, Russian-language and English-language discourse essentially boils down to using SVD first to identify the *similarities*, i.e., what topics describe both the English and Russian subcorpora within one overall corpus. The *differences* are what is left over.

One way we can empirically validate our approach, therefore, is to look at how well it does at matching up documents in different languages that are, in fact, on the same topics. And we can construct an empirical experiment where the extent to which this actually happens is objectively measurable. The experiment we perform is the same cross-validation test described elsewhere, e.g. (Chew et al., 2011); the reader can refer to that literature for the details, but we recapitulate it briefly here.

We use one parallel corpus as described above to form translation term-by-term matrix Y, and a second, *different* parallel corpus to form the term-by-document matrix X. The matrix V output by SVD will then be a concept-by-document matrix where the documents of the second parallel corpus, in different languages, are related to a single set of concepts. Since the corpus is parallel, we know *a priori* which documents are translations of one another, though this information is not available to or used by SVD. It stands to reason that a document and its translation should be similar to one another (score similarly on all topics) in the cross-language concept space. In our framework – specifically, because the SVD factorization is in an orthogonal space – 'similarity' is directly measurable by the cosine between document vectors. Therefore, we would expect to see that for any document, its translation(s) should be the most similar other documents. The approach as a whole can thus be validated by calculating accuracy as follows:

$$accuracy \approx \frac{t}{n}$$

where t is the number of documents with a translation as nearest neighbor and n the total number of documents in the second parallel corpus.

Our results, with direct comparison to prior work, are shown in Table 2. Several comments can be made on these results:

- Our approach 'does more with less' than the prior approach. It achieves significantly higher accuracy in a fraction of the processing time.
- The prior approach computes a topic space from a parallel 'training' corpus. Our approach computes the topic space directly from the 'test' corpus whether that corpus is parallel or not. The prior approach therefore *cannot* be used in the way we ultimately want – to extract cross-language topics directly from non-parallel Twitter data. But even on the task which the prior approach is good for, our approach still outperforms it.
- In addition, we tried two variants of PPMI (mentioned above) under the new approach. One is regular PPMI – calculate the pointwise mutual information between a given term and document, and set it to zero if it is negative.

The other, modified PPMI (MPPMI) never allows the weighted value for a given term, after the YX matrix multiplication, to be higher than the maximum observed for that term in X. In other words, we never allow a term to receive a greater weighting in translation than in its native language. This small modification has a significantly positive effect, as can be seen.

# SVD topics	Settings	Preprocessing run time (hrs)	Accuracy[2]
Previous approach			
300	**(Chew et al., 2011)**	43.068	0.8543
90		3.929	0.6888
Current approach			
90	PPMI, L2 norm.	0.495	0.8918
90	MPPMI, L2 norm.	**0.242**	**0.9479**

Table 2. Prior vs. current approach, same data

5 Application

Having validated the approach, we deployed it on the 206,734 Twitter posts already mentioned. We derived the X matrix from this corpus, then multilingualized it with a Y 'dictionary' matrix derived from the Bible in Russian and English. For improved results, we supplemented this 'dictionary', manually adding translation entries in the matrix for the top 100 most frequent out-of-vocabulary words, a routine process which took the first author (who knows Russian) around 1-2 hours. We then computed SVD (see section 3) using 90 latent concepts. Since no 'ground truth' exists on how documents from Twitter should align, the best validation we can do is to review the topics and verify (based on knowledge of English and Russian) that documents and terms are appropriately grouped. Top terms and posts (in both Russian and English) for sample topics include:

Topic 3 (Weighting: English .0975, Russian .1177)
Top terms: Russia, with, с (with), Россией (Russia), war, new, called, России (Russia), РФ (Russian Federation), to
Top posts:

- Russian: 'Russia news – new NATO head in Europe called for a conflict with resurgent Russia'[3]
- English: 'Could Russia REALLY go to war with NATO?'

Topic 4 (Weighting: English .0864, Russian .1317)
Top terms: summit, Poland, secretary, Warsaw, Варшаве (Warsaw), генсек (secretary), саммите (summit), Польше (Poland), саммит (summit), Порошенко (Poroshenko)
Top posts:

- Russian: 'NATO secretary will discuss in Poland the coming alliance summit in Warsaw'[4]
- English: '#Poland #Warsaw NATO summit to raise military presence in Poland, region'

It should be clear from the topics above that both documents and terms on similar topics are being appropriately grouped together (e.g. 'Russia' with its translation, etc.).

Topic 6 is mostly about internal US politics and therefore its weighting more towards the English information space is plausible:

Topic 6 (Weighting: English .1115, Russian .0941)
Top terms: на (on), Корею (Korea), blast, Korea, Клинтон (Clinton), north, Clinton, речь (speech), speech, on
Top posts (none in Russian in top 100):

- Clinton To Blast Trump On North Korea, NATO In Foreign Policy Speech. 😁😁😁
- Reuters: Clinton to blast Trump on North Korea, NATO in foreign policy speech

We found a similar weighting towards the Russian information space (Russian and English weightings were .1408 and .0775) for a topic in which 'Poroshenko'[5] featured strongly. (Petro Poroshenko, better known in Russia and Ukraine than in the West, is the Ukrainian president.)

Perhaps the most interesting point that came to our attention related to topic 11, where the words 'global' and 'strike', and Russian translations of those terms, featured among the top 10 terms for the topic. This topic had Russian and English weightings of .1152 and .0992 respectively. The authors

[2] An increase ≥ 0.018 in accuracy is always significant at p = 0.001, based on a chi-test.
[3] Russian: 'Новости России - Новый главком НАТО в Европе призвал к борьбе с возрождающейся Россией'
[4] Russian: 'Генсек НАТО обсудит в Польше предстоящий саммит альянса в Варшаве'
[5] Russian: 'Порошенко'

made the connection to NATO's 'global strike' program, which is intended to be a rapid defense capability. However, the top 100 posts scoring highly on this topic, 99 of which were in Russian, include posts such as '#News. An announcement was made in the [Russian] Federation Council of a 'global strike' by NATO on Russia /#Russia'[6]. On closer inspection it turned out that our corpus contained 2,541 Russian posts including the words 'global' and 'strike', but only 78 English posts with those words.

Our multilingual topic-extraction framework thus led us to an unexpected finding: in the Russian information space, it appeared that there was significant interest (whether from trolls, bots, or genuine users, we do not know, although that is perhaps beside the point) in 'global strike'. Further, it appears interest has been stoked (again, whether mischievously or not, we do not know) by Russians' misapprehension as to what 'global strike' actually is – development of a capability, not preparation for an actual strike on Russia. We think that it could be highly useful for policymakers and those in the diplomatic and intelligence community to have early warning of this kind of misapprehension (or disinformation?) to allow it to be effectively countered.

Further, relating this finding to sentiment and 'master narratives', an analyst familiar with the 'Fortress Russia' master narrative (Rosefielde, 2006) (that the West generally and NATO in particular is 'out to get' Russia) might explain the strong focus of attention among the Russian-speaking community on the 'global strike' topic as caused by consternation feeding from that master narrative.

To reiterate, we did not have to 'look for' this finding or know a priori that the 'signal' was there in the data. The finding simply 'fell out' from our data-driven analysis, because the number of related Twitter posts made it a significant enough pattern in the data. Our topic-extraction tool helps the analyst essentially by sorting the large 'haystack' into subcategories that make sense, allowing the analyst to get a quick sense of what is in the data, focusing data exploration on what makes the Russian part of the data *different* from the English part. This quick sort provides a defensible basis for further analysis.

[6] Russian: '#Новости. В Совфеде заявили о подготовке «глобального удара» НАТО по России /#Россия'

6 Conclusion

We have proposed a novel method which sidesteps what we see as the problems of sentiment analysis and 'master narratives' analysis, but still provides a computational solution that helps address what we believe is the over-arching social science problem with social media text: how to assess differences between the information spaces of two subsets of a corpus, with particular emphasis on differences that are most prominent in the *data* rather than in the *mind of the analyst*. We think that our results show that the sort of differences that fall out of the analysis are likely to include differences of sentiment and use of narratives – but may also include other differences that are still of interest to an analyst.

Technically, our approach involves an elegant modification to standard Latent Semantic Analysis. Our approach enables LSA to deal with multilingual input, ensuring that topics are cross-lingual to the extent possible, and projecting documents in different languages into a single multilingual topic space.

The results we have obtained demonstrate empirically that our proposed approach not only has promise as a useful analytical tool for social modeling, but also significantly outperforms previous similar state-of-the-art by simultaneously increasing both accuracy and efficiency, even with non-parallel and noisy corpora like Twitter data. Topics are cross-lingual as expected, and documents and terms that are topically related but in different languages are successfully grouped together. But more importantly, our approach (unlike prior approaches) can focus in on the topic space of a multilingual corpus whether that corpus is parallel or not – a key advantage, since most corpora that will be of interest for social modeling are not parallel.

Finally, our proposed linear-algebraic integration of machine translation and LSA does not have a precedent that we are aware of. It is the creation of this new model from information retrieval, specifically, that has opened the door for us to rethink sentiment analysis and master narratives approaches, and allow our analytics to tell us: what are the similarities, differences and patterns actually *in the data* – not those we superimpose on the data from elsewhere. This is the type of question, in our view, data

analytics is best-placed to answer with social media and other big data. Our results demonstrate ways in which this approach can avoid some of the limitations and pitfalls of sentiment and master narratives analysis, drawing useful analytical information out of big data – and turning the data's multilinguality into an asset instead of an obstacle.

7 Acknowledgement

This work was partially funded by the Office of Naval Research under contract number N00014-16-P-3020.

8 References

Beasley, A. & Mason, W. (2015). Emotional States vs. Emotional Words in Social Media. WebSci 2015 (Proceedings of the ACM Web Science Conference), article no. 31.

Brown, P., Della Pietra, V., Della Pietra, S., & Mercer, R. (1994). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), 263-311.

Center for Computational Analysis of Social and Organizational Systems. (2016). Multilingual Twitter Sentiment Analysis. Retrieved on July 27, 2016 from http://www.casos.cs.cmu.edu/projects/projects/mltsa.php.

Chew, P. (2015). 'Linguistics-Lite' Topic Extraction from Multilingual Social Media Data. *Social Computing, Behavioral-Cultural Modeling, and Prediction*. *Lecture Notes in Computer Science*, 9021(2015), 276-282.

Chew, P., Bader, B., Helmreich, S., Abdelali, A., & Verzi, S. (2011). An Information-Theoretic, Vector-Space-Model Approach to Cross Language Information Retrieval. *Natural Language Engineering*, 17(1), 37-70.

Dumais, S. (1991). Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers*, 23(2), 229-236.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Golub, G., & Van Loan, C. (1996). *Matrix Computations*. 3rd edition, Baltimore, MD: Johns Hopkins University Press.

Halverson, J., Corman, S., & Goodall, H. (2011). *Master Narratives of Islamist Extremism*. New York, NY: Macmillan.

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. *COLING '04 (Proceedings of the 20th International Conference on Computational Linguistics)*, 1367–1373.

Landauer, T., & Littman, M. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, 31-38.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, 79-86.

Rosefielde, S. (2006). Turmoil in the Kremlin: Sputtering toward Fortress Russia. *Problems of Post-Communism*, 53(5), 42-50. DOI: 10.2753/PPC1075-8216530504

Sindhwani, V. & Melville, P. (2008). Document-word co-regularization for semi-supervised sentiment analysis. *Proceedings of the 8th IEEE International Conference on Data Mining*, 1025–1030, Washington, DC: IEEE Computer Society.

Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: Wiley.