Discovering word associations in news media via feature selection and sparse classification

Brian Gawalt
Department of EECS
University of California,
Berkeley, CA 94720
gawalt@eecs.berkeley.edu

Jinzhu Jia
Department of Statistics
University of California,
Berkeley, CA 94720
jzjia@stat.berkeley.edu

Luke Miratrix
Department of Statistics
University of California,
Berkeley, CA 94720
luke@stat.berkeley.edu

Laurent El Ghaoui
Department of EECS
University of California,
Berkeley, CA 94720
elghaoui@eecs.berkeley.edu

Bin Yu
Departments of Statistics and EECS
University of California,
Berkeley, CA 94720
binyu@stat.berkeley.edu

Sophie Clavier
Dept. of International Relations
San Francisco State University
San Francisco, CA 94132
sclavier@sfsu.edu

ABSTRACT

Our perceptions of the world are largely shaped by news media. Understanding how media portray certain words and terms is a critical step towards assessing media's influence on those perceptions.

In this paper we analyze the "image" of a given query word in a given corpus of text news by producing a short list of other words with which this query is strongly associated. We use a number of feature selection schemes for text classification to help in this task. We apply these classification techniques using indicators of the query word's appearance in each document used as the document "labels" and the indicators for all other words as document predictors/features. The features selected by any scheme is then considered the list of words comprising the query word's "image".

To be easily understandable, a list should be extremely short with respect to the dictionary of terms present in the corpus. The approach thus requires aggressive feature (word) selection in order to single out at most a few tens of terms in a universe of hundreds of thousands or more. In addition, a word imaging scheme should scale well with the size of data (number and size of documents, size of dictionary).

We produce one scheme for feature selection through a sparse classification model. A standard classification algorithm assigns one weight per term in the predictor dictionary, in order to maximize the capacity to successfully predict the labels of document units. By imposing a sparsity constraint on the weight vector, we single out the few words that are most able to predict the presence or absence of a query word in any document. This paper compares this and several other schemes that are potentially well suited to the task of word imaging, each method presenting a different manner of feature selection.

We present two evaluations of these schemes. One evaluates the predictive classification performance of a logistic regression model trained over the corpus using only a scheme's selected features. The other is based on the judgement of human readers: a pair of word lists generated by different schemes operating on identical queries are presented to a human subject alongside a trio of document units (paragraphs) containing the query word. This subject then chooses the list which in his/her estimation is the best summary of the document units.

We apply these schemes to study the images of frequently covered countries and regions in recent news articles from the International section of the New York Times. Our preliminary experiments indicate that, while most methods perform similarly well on our data in terms of predictive performance, human-based evaluations appear to favor features selected by training of sparse logistic regression, a penalized variant of logistic regression that encourages sparse classifiers. This indicates that classification metrics based on pure predictive performance, while useful as a indicator for preselecting algorithms, are not enough to predict human assessment of word association algorithms.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Information filtering*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology*

General Terms

Algorithms, Measurement, Performance, Experimentation, Human Factors

Keywords

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

Sparsity, Evaluation, News Media, Text Classification, Logistic Regression, Regularization.

1. INTRODUCTION

Motivating application: analysis of news. Progress in technology allows the public daily access to an unprecedented volume of news, coming from various sources. However , given this volume, significance and meaning of these reports can be hard to assess and decipher. Can machine learning help? In turn, can this field of news media analysis modify and inform progress in machine learning?

Text classification is a vibrant field [17, 18] that has been extensively used for news data. It can help to categorize documents [17, 5, 10], to provide sentiment analysis [12] and opinion mining on articles [15], and to predict future market trends. It appears that relatively little has been done to connect the extensive technological progress in the area of text classification to issues that are of concern to the social scientist, such as how the media does or does not influence our perception of the world.

There is a long academic tradition in humanities and social sciences scholarship of extracting quantitative data from manual classification methods and then qualitatively assessing the significance of usage patterns within the resulting categories and category schemes [2, 7]. Many of these media studies usually make use of simple word frequency or co-occurence counts. Recent work such as [6], in which the authors seek to quantify media slant, calls into play more elaborate statistical methods. There does seem to be an opportunity for a strong interplay between text classification methods, as presented in the machine learning literature, and qualitative approaches to discourse analysis. A call for such an effort (within the context of literary discourse) has been made by Moretti in [13].

Studies of word usage in the media studies have often focused on the portrayal of international issues in domestic or international media [14]. The need for such analyses is indeed acute for news related to foreign policy. Jervis in 1976 [8] opened foreign policy analysis to include the role of perceptions and misperceptions in international politics, especially in terms of security. In addition, many media studies have clearly identified the mainstream media as the primary source of information about world affairs. Entman's book, *Projections of Power* [3], argues that the media not only frames the agenda, but is part and parcel of the exercise of power, specifically where the predominant frame is transmitted from the executive to other elites, then to the media and finally to the public. We agree with Thompson [19] that "media presentation is a crucial determinant of the public perception of international politics," whether or not this frame-making is intentional. Therefore any systematic and easy-to-use tool that can help in understanding *how* issues, events and policies are presented in the media, has a fundamental role to play in social sciences. One such tool could provide analysis of word associations in a given corpus, allowing one to better understand how concepts are linked to shape our perceptions.

Our focus: word imaging. Our focus is the potential use of text classification in understanding the "image" of a certain word in a given news stories corpora. By image, we simply refer to the words in the dictionary that, in the corpus and within the time window under consideration, are associated in some statistical sense with the word under study (we will refer to this as the "query" word). We believe that such images could provide a great service to researchers in media studies.

Our approach: feature selection. There are many ways to define association, from co-occurence of words within (say) a paragraph to more sophisticated methods. A first challenge in the imaging problem is thus to choose a meaningful method to specify and quantify association between terms. Our methodology rests on feature selection, which roughly refers to a set of algorithms that perform efficient reduction of the number of features considered by a classification model while maintaining satisfactorily minimal classification error rates.

Based on a given query term, we separate the news corpus document units[1] into two classes: those that contain the query and those that do not. We then run these units through the paces of a feature selection scheme, resulting in an assignment of scheme-distinct weights/scores to each term appearing in any unit. Only the highest scoring terms are retained; for these purposes, only the top fifteen or so are kept. In the estimation of the particular feature selection scheme, the resulting short list is an image of the query.

In text applications, the number of features, n, is typically very large[2]. We are aim to provide a web-based news analysis application for social science researchers in the near future, which by its nature requires an extremely short calculation times. Hence, a satisfactory feature selection scheme should scale well with problem size, capping the complexity of admissible schemes. We have not yet included sophisticated natural language processing techniques[11], leaving this to future research.

Example. To illustrate our feature selection approach, let us show how it can be used to visualize the history of a term's image. We have applied sparse logistic regression as a feature selection scheme (see Section 2.3 below) to study of the term "microsoft" in all the New York Times headlines between January 1981 and December 2006. (Each headline is a document unit.) We have used the "BBR" algorithm implemented in [5] on a sliding window of one year's worth of data taken at yearly increments. We looked at one year's headlines at a time and trained a sparse logistic regression classifier, tuned to produce fewer than 10 nonzero-weighted

[1]These document units are free to be the full article texts, single paragraph texts, the headlines – an experimenter may choose how to atomize the corpus based on the degree of inclusivity or specificity appropriate for the media phenomena being investigated.

[2]A reasonably sized corpus can easily contain tens of thousands of unique words. The number of digrams, distinct pairs of words which ever appeared consecutively, can extend into the millions.

words for any year's documents, to distinguish headlines containing the term "microsoft" from those not containing that term, repeating this for year after year. Thus, for each year, we have obtained a short list of terms, and associated weights, that are active predictors of the appearance of the query term in any headline. This results in a matrix of weights, each column corresponding to a year, and each row to a word that ever appeared in a "microsoft" image. If we arrange the rows of the matrix so that words are listed in order of appearance, the matrix will have a staircase pattern. Tracing any row reveals the corresponding term blinking in and out over time as a salient descriptor of "microsoft".

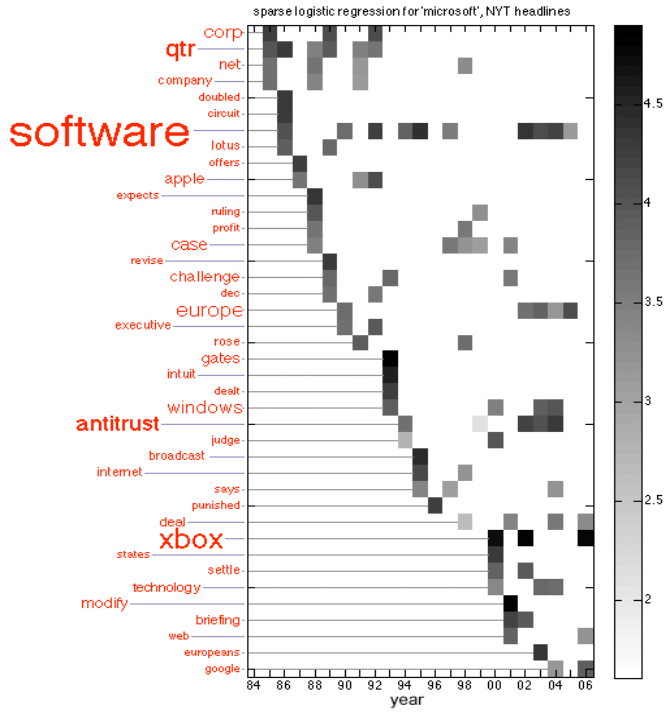

Figure 1: Study of the term "microsoft" within New York Times headlines: matrix of sparse logistic regression weights with corresponding high-weight terms highlighted.

One visualization of the matrix of weights is shown in Figure 1. In the figure, each little rectangle indicates the presence of that row's word as an important feature for that column's year. The darkness of the rectangle indicates its relative weight to the other words selected for that year. The vertical axis corresponds to the terms, shown by order of appearance; thus, the plot shows a staircase pattern, where we have emphasized the font of terms with a large total sum of absolute weights over time. Table 1 provides a list of the top 30 of such words.

The list in Table 1 appears to provide an accurate summary of Microsoft, with the top prize going to "software" (the long-term focus of the company) and "xbox" (its most recent best-selling product). Figure 1 goes further, in providing a story of the evolution of the company that is consistent with common knowledge. The initial terms refer to a high-growth corporation (with terms like "company", "net", "profit", "doubled", and "qtr", a common business-news abbreviation for

1	software	11	apple	21	executive
2	xbox	12	challenge	22	rose
3	qtr	13	modify	23	internet
4	antitrust	14	briefing	24	broadcast
5	europe	15	technology	25	ruling
6	corp	16	deal	26	says
7	windows	17	settle	27	expects
8	case	18	intuit	28	europeans
9	gates	19	lotus	29	dec
10	net	20	company	30	profit

Table 1: Most important words found by sparse logistic regression analysis of the term "microsoft" in The New York Times headlines, ranked by sum of absolute regression coefficients over time.

"quarter"). The list of words then visits terms related to products, from "lotus" to "windows" to "xbox". Another important topic involves legal terms ("case", "judge", "settle"), with a reference to the famous anti-trust case in Europe. More recently, the terms reflect the growing importance of the Internet ("web") and media ("broadcast"). Throughout, the names of important related companies are mentioned: "lotus", "apple", "google".

Reading the plot vertically gives the main topics for a particular year. For example, the year 2002-2003 has "anti-trust", "europe", and "software". The plot also allows one to pinpoint terms that frequently recur in the news over a long stretch of time (e.g., "software").

This particular example is encouraging in that this and other feature selection schemes could be useful in providing a quantitative, time-consistent, common-sense summary of a widely cited topic. Obviously the question arises as to what scheme should be used.

Contributions. In this paper, our aim is to evaluate several schemes over full article text drawn paragraph by paragraph from The New York Times International section. (The schemes are potentially well suited to a real-time imaging task.) We use two styles of evaluation: predictive classification performance and human evaluation based on a rigorous protocol of comparison. We find that even though the predictive performances of many schemes are similar, human-based evaluation seems to favor sparse logistic regression as a feature selection scheme. This implies that predictive performance alone is not enough to choose algorithms for the word imaging task, and further research is needed to better understand "what humans want" in terms of word images and to see whether sparse logistic regression can serve as an automated method for the word imaging task more generally.

Our paper is organized as follows. We provide an overview of the schemes used in Section 2, including details on preprocessing the text data. We describe our statistical metrics based on predictive performance in Section 3. Section 4 is devoted to our human evaluation protocol and the results reported by five volunteer readers. In both these sections, we provide results pertaining to the image of various oft-

cited countries and regions in the international section of The New York Times between December 2008 and October 2009.

2. FEATURE SELECTION SCHEMES

The five schemes we tested fall under two approaches. Four algorithms use independent feature models in their selection: co-occurence count (COOC), Delta-TFIDF (D-TFIDF), Binomial Normal Separation (BNS),and a threshold of p-values as calculated assuming a χ^2 log-likelihood of appearance rates (CHI). A fifth feature selection method, sparse logistic regression (L1LR), does not rely on independence assumptions, and instead uses a l_1-penalized variant of logistic regression to select features. (By abandoning these assumptions, a cost is borne in terms of increased computational complexity.) We detail these schemes in turn after describing how we transformed raw text data into a numerical form.

2.1 Pre-processing

Corpus. Our data are a series of news articles from the International section of the New York Times, as syndicated on their RSS feed. Publication dates run from December 15, 2008, to October 18, 2009. We stripped the corpus of capitalization, reverting all characters to lower case. Punctuation was also scrubbed, with marks deleted and their neighboring characters joined. White space was maintained (For example, the plaintext "Asian-American Studies" becomes "asianamerican studies".) From here, the text was vectorized. In our analyses, we only considered single-word terms (no digrams or trigrams), and we considered each paragraph to be a single document unit.

Bag of words. To extract the statistical structure of a corpus, the news data must first be somehow enumerated. We used a "bag-of-words" approach, where each paragraph of the corpus is represented by a vector whose dimensionality includes one element for each distinct word. The j-th element for vector i is then set to the number of times word j appears in the i-th document unit. Our news corpus is comprised of $79,494$ distinct words (our dictionary) used across 109,686 paragraphs, leading to a data matrix $X \in \mathbf{R}^{m\times n}$, with $m = 109,686$ rows (number of data points) and $n = 79,494$ columns (dimension of feature space, that is, dictionary size). As most paragraphs have a word count under sixty, less than 0.05 percent of elements of this matrix are nonzero.

Document labels. We sought to distinguish between paragraphs containing a given query word and those that did not. Let $q \in \{1,\ldots,n\}$ be the index of the query word. We labeled each paragraph i as a positive example if the query word appears in it at least once, and negative otherwise, that is:

$$y_i(q) = \begin{cases} +1 & \text{if } X_{iq} > 0, \\ -1 & \text{else,} \end{cases} \quad i = 1,\ldots,m.$$

The number of positive examples for our several experiments varied between fifty and two-thousand (representing between roughly 0.05% and 2% of the document units). In our experiments, we always remove the q-th column in the data matrix, which corresponds to the query word.

Stop words. In many cases, words may be deemend intrinsically uninteresting. Terms such as "in", "with", "and", "but", "the", etc., carry little-to-no descriptive weight. We can state a priori that they have no place in a word image. Dropping them from the matrix before processing the data costs little, helps decrease runtime, and ensures more descriptive images. However, this process is not riskless: while the word "said" is typically used as a neutral linking verb with little connotative value, its proper noun heteronym "Said" (perhaps as in Edward Said, the literary theorist), is indeed informative. We have used a limited list of 300 words, available at our web site[3]. These 300 words were removed from the dataset.

Stemming. Many distinct words share meaning: verbs can describe an identical action but vary by tense, a noun can be another noun's plural, etc. As with stop words, it can be helpful to reduce the size of the overall dictionary by mapping words with shared roots into a common feature. This process of stemming is common in many applications. We have declined to stem our dataset. The same risk of lost information as in stop words above applies even more severely here. For example, a stemmer might be expected to consider "iraq" and "iraqis" equivalent, but the connotation of an image that focuses on a nation's individual citizens as opposed to one focused on the nation as a whole is an important distinction for our purposes. We cannot tune stemming with the same precision allowed by the stop-word list above. Note that some authors recommend stemming in text classification [18], while others warn of a potential loss of predictive performance [17].

After these steps, the dataset is ready for statistical analysis. The algorithms we used are all based on a first step where feature selection is performed. Then a standard logistic regression algorithm, described next, is applied to assign weights to the selected features.

2.2 Independent-model feature selection methods

Our words have been indexed by set $J = \{1,2,...,n\}$ and document units (paragraphs) by the set $I = \{1,2,...,m\}$. Given a query q, these document units have been perfectly partitioned into two subsets, $I^+(q) = \{i \in I | y_i(q) = +1\}$, of cardinality $\#I^+(q) = m_q^+$, and $I^-(q) = \{i \in I | y_i(q) = -1\}$, of cardinality $\#I^-(q) = m_q^-$. (Note, $m = m_q^+ + m_q^-$ for all q). We sought by each scheme a subset $K \subseteq J$ with cardinality as close as possible to a given target k. The schemes are summarized below.

Co-occurence (COOC). For each word $j \in J$, we compute the COOC score $c_j^+(q) = \sum_{i\in I^+(q)} X_{ij}$. Let $\bar{c}_k(q)$ be the $k+1$th highest value found in vector $c^+(q) = \{c_j^+(q)\}$. By this method, we build $K_c^k(q) = \{j \in J \ : \ c_j^+(q) > \bar{c}_k(q)\}$.

[3] www.eecs.berkeley.edu/~gawalt/MIR2010/

We have picked by this method the k non-stop-words which most frequently appear in paragraphs in which our query also appears.

Delta TF-IDF (D-TFIDF). The Delta TF-IDF method (D-TFIDF for short) [12] uses a variant of the well-known Term Frequency, Inverse Document Frequency (TF-IDF) vectorization of text documents.

Having established $c^+(q)$ above, we calculate for each word $j \in J$ the count of positive and negative paragraphs with that word. Namely, let $d_j^+(q) := \#\{i \in I^+(q) : X_{ij} > 0\}$ and $d_j^-(q) := \#\{i \in I^-(q) : X_{ij} > 0\}$ Note that $d_j^+(q)/m_q^+$ is the percent of times word j appears at least once in the positive examples. Similarly for $d_j^-(q)$.

We use these values to produce the D-TFIDF score:

$$\delta_j(q) = c_j^+(q) \log\left(\frac{m_q^+}{d_j^+(q)} \frac{d_j^-(q)}{m_q^-}\right), \quad j = 1, \ldots, n.$$

Let $\bar{\delta}_k(q)$ be the $(k+1)$-th highest value found among the magnitude of these values $|\delta(q)|$. We build $K_\delta^k(q) = \{j \in J : |\delta_j(q)| > \bar{\delta}_k(q)\}$. This method selects by a combination of seeking words that appear commonly alongside our query term, with added sophistication to penalize those words that co-occur too often in the positive examples (perhaps an indication of what is effectively, for this query word, a stop word) and rewarding those that appear rarely in our negative example paragraphs.

Bi-normal separation (BNS). The bi-normal separation (BNS for short) method has been proposed in [4].

We take vectors $d^+(q)$ and $d^-(q)$ as above. For each word $j \in J$, we compute the BNS score $b_j(q) = \Phi^{-1}\left(\frac{d_j^+(q)}{m_q^+}\right) - \Phi^{-1}\left(\frac{d_j^-(q)}{m_q^-}\right)$, where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution. Let $\bar{b}_k(q)$ be the $(k+1)$-th highest value found among the magnitude of these values $|b_j(q)|$. We build $K_b^k(q) = \{j \in J : |b_j(q)| > \bar{b}_k(q)\}$. This method selects words with divergence between rates of appearance in each paragraph class using an underlying normal model for appearance rates. This allows for greater distinction between tail and modal behavior; extremely rare or common words are assessed by a different standard than words that appear about half the time.

χ^2 log likelihood (CHI). Multiple testing problems require identifying the significance of numerous hypotheses inferred from common data simultaneously. We take a hypothesis for each word $j \in J$: that there exists a significant difference in the appearance rate of word j between the two document classes. A p-value for each hypothesis can be calculated, and we assumed the log-likelihood ratios used in these calculations approximately follow χ^2 distributions. The ranked p-values of each hypothesis are used to judge these differences and their significance. The selection of a threshold for these p-values, accepting only those hypotheses whose values surpass the threshold, can lead to control of many types of error rate, such as the false discovery rate or family wise error rate [1, 16]. We have adapted this method to the feature selection problem.

We again take $d^+(q)$ and $d^-(q)$ as above. For each word $j \in J$, we compute the log-likelihood ratio score $f_j(q)$:

$$\begin{aligned} f_j(q) = \; & d_j^+(q) \log\left(\frac{d_j^+(q)}{m_q^+}\right) + \\ & [m_q^+ - d_j^+(q)] \log\left(1 - \frac{d_j^+(q)}{m_q^+}\right) + \\ & d_j^-(q) \log\left(\frac{d_j^-(q)}{m_q^-}\right) + \\ & [m_q^- - d_j^-(q)] \log\left(1 - \frac{d_j^-(q)}{m_q^-}\right) - \\ & [d_j^+ + d_j^-] \log\left(\frac{d_j^+ + d_j^-}{m}\right) - \\ & [m - d_j^+(q) - d_j^-(q)] \log\left(1 - \frac{d_j^+(q) + d_j^-(q)}{m}\right) \end{aligned}$$

Let $\bar{f}_k(q)$ be the $(k+1)$-th highest value found among the magnitude of these values $|f|$. We build the set of features for this scheme $K_\chi^k = \{j \in J : |f_j(q)| > \bar{f}_k(q)\}$.

Breaking ties. In certain cases, there may be a tie for the k-th highest score under any of the above schemes. Often this can be avoid by repeatedly executing the scheme on a subset of the training data, allowing scores to accumulate for words across each iteration. We reran each scheme for 10 iterations, holding out a randomly selected 10% of the training data each time. This has the added benefit of promoting stability in the word choices: the effect of outlier paragraphs on word scores can be muted in this way. Should any ties remain in the cumulative scores, the wordlist was padded out to a length of k by randomly and uniformly drawing from among the words tied for k-th place.

Assumptions of feature independence. The appeal of the above schemes lies in their scalability. The order of computational complexity is linear in the number of distinct words and documents. However, this scalability results from an underlying assumption of independence between the appearances of words across documents. Below, we investigate applications of a more computationally intensive scheme which may take advantage of dependence between word use patterns, test whether this approach leads to better word images, and consider whether the image improvement is worth the increased computation costs.

2.3 Logistic regression

Logistic regression is a classical classification method based on a generalized linear model [5]. For each query q, Take data points $x_i \in \mathbf{R}^n$ and associated labels $y_i(q) \in \{-1, 1\}$, $i = 1, \ldots, m$. The logistic regression model is based on the following expression for the conditional probabilities:

$$P(y_i(q) = 1|x_i) = \frac{1}{1 + \exp(-x_i^T\beta - \gamma)},$$

where $\beta \in \mathbf{R}^n$ is the vector of regression coefficients in the model, also referred to as "weights", and $\gamma \in \mathbf{R}$ is an intercept. An estimate of the vector β can be obtained by solving the corresponding maximum (log-)likelihood problem, which can be expressed as

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta,\gamma} \; -\mathcal{L}(\beta, \gamma), \tag{1}$$

where

$$\mathcal{L}(\beta, \gamma) := -\sum_{i=1}^{m} \log\Big(1 + \exp(-y_i(x_i^T\beta + \gamma))\Big)$$

is the log-likelihood function. Logistic regression has been widely used in data mining and text classification [5].

Sparse logistic regression (L1LR). Sparse logistic regression (L1LR for short) [5] allows for simultaneously performing feature selection and model fitting, via the introduction of an l_1-norm penalty to the maximum-likelihood problem:

$$(\hat{\beta}(\lambda), \hat{\gamma}(\lambda)) := \arg\min_{\beta,\gamma} \; -\mathcal{L}(\beta, \gamma) + \lambda \sum_j |\beta_j|, \tag{2}$$

where $\lambda > 0$ is a penalty parameter. The presence of the l_1-norm encourages many components of the estimated vector β to be zero, an effect that becomes more pronounced as $\lambda \to +\infty$. The lower bound of $\lambda > 0$ and the upper bound for λ provided by [9] (beyond which all weights of the classifier are zero) allows for fast binary line search to obtain a λ such that only k nonzero elements remain in vector $\hat{\beta}(\lambda)$. We used the efficient BBR software described in [5]. We encountered no need for a tie-breaking mechanism alongside this process. The search for a λ to produce k nonzero elements was completed reliably and quickly for all topics considered; whatever ties might be occurring are handled by the BBR implementation without a need for additional preprocessing from us. According to our experiments, for a fixed λ, the BBR algorithm takes about 15 seconds for a problem with 87,743 observations and 72,443 words[4] when conducted on a typical laptop computer[5]. (Computation times for the independent-model schemes are too small to measure.) While L1LR is more computationally intensive than the independent-feature schemes described above, the speed and efficiency of present implementations and hardware are together sufficient to admit L1LR's use in analyzing news data.

In our approach, we used L1LR purely as a feature selection mechanism. As with the four previous approaches, once the features are identified this way, we resorted to (unpenalized) logistic regression on the selected features to obtain the classifier weights.

[4]This reduction in dictionary size reflects not only the removal of stop words in the preprocessing step, but also those which never appear in the training set.

[5]The specific hardware had CPU speeds near 2.7GHz, and 64MB of memory was used by the software.

3. PREDICTIVE PERFORMANCE EVALUATION

Although we do not focus on text classification predictive performance as an end in itself, we would like to explore whether classification performance could be used as a proxy to evaluate which method gives "better" results for this word imaging task. (We elaborate on what "better" means in Section 4 on Human Evaluation.)

Train-test split. As is standard procedure in machine learning, we have divided the dataset into two partitions. The documents from the larger partition are used to train a classifier according to a particular model or algorithm, and the predictions this classifier yields on the smaller partition's documents are compared to their true labels. Given a query q, we executed a random split while ensuring that there are four training documents for every test document and that the proportion of positive examples to negative examples are equivalent in each partition. Rows marked for testing are removed from data set $[X, y(q)]$ and stacked in test set $[X^{\text{test}}, y^{\text{test}}(q)]$.

Procedure. We selected a list of 47 query terms from the list of frequently-cited countries and regions in our corpus[6]. For each training set $[X, y(q)]$ associated with each of 47 query words found in the news corpus, and for a cardinality target k held constant across all schemes, we established five logistic regression models. k varied across queries from 10 to 17 terms.

We first utilized one of the above feature selection methods: COOC, D-TFIDF, BNS, CHI, and L1LR. Each in turn was used to create its own feature set K', from which a particular input matrix could be crafted $X' = \{x_{ij} : i \in I, j \in K'\}$. This matrix, combined with label vector $y(q)$, was used to produce a logistic regression model (now of dimensionality k, thanks to the pruning of aggressive feature selection), leading to a vector of coefficients β' and intercept γ'.

Each of the trained logistic models is then used to generate predictions $\hat{y}(q)$ based on X^{test}. Given an input test vector x^{new} and a logistic model with parameter (β, γ), a probability

$$P(x^{\text{new}} \mid \beta, \gamma) = \frac{1}{1 + \exp(-\beta^T x^{\text{new}} - \gamma)}$$

can be calculated. For a given threshold $\bar{p}$, we then predict the new label as follows:

$$\hat{y}^{\text{new}}(q) = \begin{cases} +1 & \text{if } P(x^{\text{new}} \mid \beta, \gamma) > \bar{p}, \\ -1 & \text{otherwise.} \end{cases}$$

In our experiments, we used $\bar{p} = 0.5$ unless otherwise noted.

[6]This list of queries, as well as the words generated from the training data by each scheme, is provided on our web site, `www.eecs.berkeley.edu/~gawalt/MIR2010/`.

The performance of the prediction as compared to the known, true values in $y^{\text{test}}(q)$ was evaluated according to four well-known scores: precision, recall, F1 (for a given $\bar{p}$) and Area-Under-Curve (AUC, a metric which considers all $\bar{p} \in [0, 1]$). Precision measures the ratio of the number of correct positive predictions $\#\{i : \hat{y}_i(q) = 1,\ y_i^{\text{test}}(q) = 1\}$ to the number of positive predictions $\#\{i : \hat{y}_i(q) = 1\}$; recall measures the ratio of number of correct positive predictions to the number of positive examples $\#\{i : y_i^{\text{test}}(q) = 1\}$. As there traditionally exists a tradeoff between these two measures, the measure $F1$, which is the harmonic mean of precision P and recall R, $F1 = \frac{2PR}{P+R}$, can be used as a summarization of the two.

The AUC score requires a sweep of the $\bar{p}$ parameter from 0 to 1, establishing true positive rate:

$$\text{TPR} = \frac{\#\{i : \hat{y}_i^{\text{test}}(q) = 1,\ y_i^{\text{test}}(q) = 1\}}{\#\{i : y_i^{\text{test}}(q) = 1\}}$$

and false positive rate:

$$\text{FPR} = \frac{\#\{i : \hat{y}_i^{\text{test}}(q) = 1,\ y_i^{\text{test}}(q) = -1\}}{\#\{i : y_i^{\text{test}}(q) = -1\}}$$

Plotting TPR against FPR for each $\bar{p} \in [0, 1]$ provides the receiver operator characteristic for the classification model, and the area under this curve (AUC) is a metric of the model's fitness to the data.

Results. The boxplots for the precision, recall, $F1$ and AUC metrics as calculated for the 47 queries can be found in Figure 2. Tables 2, 3, 4, and 5 display the results of one-sided paired T-tests. These tests establish a p-value for the metric's hypothesis, "Given results for all queries, scheme a scores higher on this metric than scheme b." We can see a pattern in these tables and figures: classification based on features selected by COOC or BNS underperforms compared to classification based on features selected by the schemes L1LR, D-TFIDF, and CHI. These three "winners" are largely indistinguishable from each other in their effects on predictive performance, as are the two "losers". (COOC does outperform BNS on the metric of precision.)

	L1LR	CHI	DTF	COOC	BNS
L1LR		0.29	0.34	0.10	**0.03**
CHI	0.71		0.62	0.20	**0.03**
DTF	0.66	0.38		0.16	**0.03**
COOC	0.90	0.80	0.84		0.14
BNS	0.97	0.97	0.97	0.86	

Table 2: Precision. p-values from a one-sided paired T-test to compare the five different schemes on the metric of classifier precision, addressing the hypothesis, "Does the scheme indicated by the column outperform the scheme indicated by the row?" Significant comparisons ($p < 0.05$) are high-lighted.

4. HUMAN EVALUATION

The Experiment. We have seen the performance of the five schemes in terms of classification error. Ideally, for a given query, the main features selected for classification also capture an inherent aspect of the query in a given data set. This potential connection is a theory that we aimed to substantiate with a human validation experiment. In our experiment, human readers were given a survey of 60 questions. For each question, the subject read three random paragraphs about a query and selected which of four word lists (as generated by schemes above) best captured the image of those paragraphs. They then were asked to identify the common query itself as a way to validate the paragraphs selected.

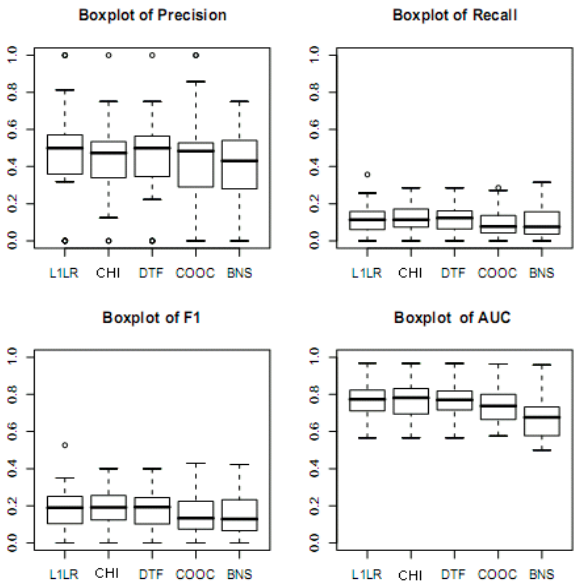

Figure 2: Statistical Evaluation

	L1LR	CHI	DTF	COOC	BNS
L1LR		0.96	0.87	**0.00**	**0.04**
CHI	0.04		0.11	**0.00**	**0.01**
DTF	0.13	0.89		**0.00**	**0.01**
COOC	1.00	1.00	1.00		0.73
BNS	0.96	0.99	0.99	0.27	

Table 3: Recall. p-values from a one-sided paired T-test to compare the five different schemes on the metric of classifier recall, addressing the hypothesis, "Does the scheme indicated by the column outperform the scheme indicated by the row?" Significant comparisons ($p < 0.05$) are high-lighted.

For each human subject, and given the set of queries Q, we repeated the following steps 60 times:

1. Select two schemes a and b. Select a query $q \in Q$. Select a "decoy", $r \in Q$, $r \neq q$. Select three paragraphs at random that mention the query but do not mention the decoy, i.e. select uniformly from the set $\{i : i \in I^+(q), i \notin I^+(r)\}$.

2. Show the subject a screen with the three paragraphs followed by four word lists, each of a length between 10

	L1LR	CHI	DTF	COOC	BNS
L1LR		0.91	0.81	**0.00**	**0.02**
CHI	0.09		0.17	**0.00**	**0.00**
DTF	0.19	0.83		**0.00**	**0.01**
COOC	1.00	1.00	1.00		0.54
BNS	0.98	1.00	0.99	0.46	

Table 4: $F1$. p-values from a one-sided paired T-test to compare the five different schemes on the metric of classifier $F1$, addressing the hypothesis, "Does the scheme indicated by the column outperform the scheme indicated by the row?" Significant comparisons ($p < 0.05$) are high-lighted.

	L1LR	CHI	DTF	COOC	BNS
L1LR		0.45	0.68	**0.00**	**0.00**
CHI	0.55		0.78	**0.00**	**0.00**
DTF	0.32	0.22		**0.00**	**0.00**
COOC	1.00	1.00	1.00		**0.00**
BNS	1.00	1.00	1.00	1.00	

Table 5: AUC. p-values from a one-sided paired T-test to compare the five different schemes on the metric of classifier AUC, addressing the hypothesis, "Does the scheme indicated by the column outperform the scheme indicated by the row?" Significant comparisons ($p < 0.05$) are high-lighted.

and 17 words. The word lists are (in a random order) the four word lists resulting from running scheme a over q, scheme b over q, scheme a over r, and scheme b over r.

3. Ask the reader to pick their first- and second-choice word lists, among those on display, that best capture the image of the three paragraphs.

4. On a seperate screen, ask the reader if the three paragraphs are about q, r, both, or neither.

In this way, a survey question now resembles a trial or contest between the two schemes involved. We cycled through a random permutation of the 10 possible scheme pairings 6 times to balance the head-to-head scheme comparisons. An order on the set of queries were permuted, instead of sampled with replacement, and taken as the order of the queries across the 60 questions. For each question, a decoy was drawn at random. This helped maintain balance in query-decoy pairings. In all trials, word lists were truncated to the length of the shortest word list: all lists presented at a time were the same length. Since all schemes score their features, truncation always took those features with the highest magnitude weights.

In each question, there are three categories of responses step (3) could elicit. First, the list of scheme a over q could be selected as the best; second, the list of scheme b over q could be selected as the best. In both these cases, the response presents a datum in evidence of a outperforming b or b outperforming a, respectively. Third, the reader could select either a or b over r. This outcome suggests either that a and b are of low quality, or that the paragraphs fail to establish the image of the query. This is where the response to step (4) is useful; if the reader was able to correctly identify the query but rejected either scheme's image of that query, then this is evidence that the schemes are at fault. When the reader was unable to correctly name the query based on the paragraphs, the response to (3) was dropped from the analysis.

Survey responses. Five different versions of the above survey were administered to five of the project's undergraduate assistants. In 60% of the responses, only one list was picked (there was no second-choice word list given). 27% of responses had both a first- and second-choice selected. The remaining 13% had no response at all. Table 6 shows the number of times each scheme was picked first, picked second, not picked at all ("skipped"), or lost out to a selection of a decoy list ("bad"). Questions for which the subject indicated the paragraphs were not about the query (10% of all replies), or were about both query and the decoy (6%), were not included in the below analysis.

	L1LR	COOC	DTF	CHI	BNS
first	60	50	37	15	10
second	8	12	8	3	6
skipped	21	28	48	78	54
bad	3	9	6	11	29
total	92	99	99	107	99

Table 6: Number of times each scheme was picked first, second, not at all, or picked below a decoy list.

In Table 7 we show the p-values for the hypotheses, "Scheme a was selected at a higher rate than simple random chance," comparing the rate at which a scheme was selected as a first-choice to the 1 in 4 chance it would have if selected at complete random on each question. The results suggest L1LR, COOC, and D-TFIDF score well; BNS and CHI do not. This test is a conservative measure, however, of a scheme's effectiveness at imaging. It is not necessary for this test that BNS or CHI lists be themselves of low quality, only that they rarely exhibit a quality sufficient to be picked ahead of another scheme's list. We examine the dynamics of head-to-head selection of one scheme over another in the next section.

	picked	n	$\hat{p}$	P-value
L1LR	60	92	0.65	0.000
COOC	50	99	0.51	0.000
DTF	37	99	0.37	0.004
CHI	15	107	0.14	0.998
BNS	10	99	0.10	1.000

Table 7: First column is the number of trials a scheme was picked first for valid paragraph sets. Second column is number of trials where the scheme appeared at all. $\hat{p}$ is proportion of time process was picked first. p-value is for a one-sided binomial test against $\alpha = 0.25$.

Comparing the schemes. For a given survey question, a scheme's word list over the query can be picked first, picked second, or not picked. If a question pits scheme a against scheme b, we score the contest as being for a if a's query list is picked first, for b if the b's query list is picked first, and a tie otherwise. Responses where a decoy list was picked first, instead of either query list, suggest that the quality of neither schemes' list was satisfactory.

If scheme a is superior to scheme b, evidence of this could be found two ways. First, a could be picked more often than b in the head-to-head contests between the two. Second, the rate at which a is selected over a third scheme c in a-c contests could surpass the rate at which b is selected over c in b-c contests. We combine evidence of both kinds to create an overall test statistic of scheme performance.

Each survey provides six passes through the 10 possible question pairings for our five schemes. One of these questions on a given pass is the head-to-head contest of a and b. Three questions involve pitting a vs. c, d, or e, and three questions involve pitting b vs. c, d, or e. The remaining questions concern neither a nor b.

For the head-to-head contests, the null hypothesis contending that there is no difference in imaging capability between a and b holds a 50% chance of a being picked (taking as given that one of them was picked at all). We can test this hypothesis with a binomial test with $\alpha = 0.5$. Let the resultant p-value be $p_1(a, b)$.

Under the null hypothesis there is no difference between a and b, there is common probability η of either a or b being picked over c, d, or e in those contests — the rate at which a beats any of the other three should equal the rate b beats any of the other three if the two schemes are of equal quality.[7] We test this using the (approximate) χ^2 test on the two-way table of process vs. being picked first. (As long as the data determining a's performance over the others is independent of those of b, this test is a valid construction.) Let the resultant p-value be $p_2(a, b)$.

For a pair of schemes a and b, $p_1(a, b)$ and $p_2(a, b)$ are independent of each other, as they are derived from independent sets of trials. Let our test statistic of scheme quality difference be the product of the two p-values, $B(a, b) \equiv p_1(a, b) \cdot p_2(a, b)$. If the tests behind $p_1(a, b)$ and $p_2(a, b)$ were continuous and exact, then under the null the p-values are independent random variables distributed uniformly: $p_i(a, b) \sim Uni[0, 1]$. The cumulative density function of $B(a, b)$ under the null hypothesis, and equivalently $B(a, b)$'s p-value, is then:

$$\mathbf{P}\{B \leq b\} = b(1 - \log b)$$

The smaller the measurement of $B(a, b)$, the more extreme the difference between the two processes.

Although both tests are asymptotically exact and continuous, $p_1(a, b)$ is a binomial test, which is exact and discrete and $p_2(a, b)$ is a χ^2 test on the 2×2 table of Scheme $\times$ Win-Loss count, which is approximate and asymptotically exact;[8], making our final p-values for the $B(a, b)$-statistics also approximate. We therefore verified the results with a permutation test on the distribution of $B(a, b)$ results did not substantively change.

Interpreting a significant result with this test requires a modicum of care. The overall test is built out of two bi-directional tests. If these are in different directions, and we have a significant result, then we can say that the processes *differ* but not that they are ordered. Furthermore such a situation would call into question whether we could order the processes overall, and would unveil that they perform in context. If both sub-tests are in the same direction then we can interpret it as we interpret any normal test of difference—the larger/better thing is significantly larger or better. But we must remember we are testing for difference in something potentially more complex than a single dimension, and so ordering is not necessarily well defined.

Results. Using the above test statistic, we get the results shown in Table 8. For readability, $B(a, b)$ has been log-transformed so the scale is not tiny and so higher values are more extreme. We have 10 hypothesis that are definitely related. Under a Bonferroni correction, all tests are significant except for L1LR vs COOC, COOC vs D-TFIDF, and CHI vs BNS. We can conclude that L1LR is better than all the other methods except, possibly, COOC. All sub-tests for significant $B(a, b)$ are in alignment — $p_1(a, b)$ always agreed with $p_2(a, b)$ over which scheme appeared preferable — so interpretation is more straightforward.[9]

The direction of all head-to-head comparisons produces the order: L1LR, COOC, D-TFIDF, CHI, BNS. All pairwise comparisons follow this ordering. Although the top three are not entirely separated, given our original hypothesis that L1LR would produce better lists due to its consideration of interfeature collinearity, the above data do suggest its superiority. Furthermore, the p-value for the comparison of L1LR to COOC went down to 0.04 under the more exact permutation test.

Our experiment asked readers to sort out four objects of moderate complexity in relation to groups of three small blocks of text. We extend their assessment of these to the entire corpus of data. Our conclusions are still dependent on a large number of conditions and assumptions. For example, we need to verify more substantively whether the cognitive load of the readers' task is reasonable; we should certainly assess how long the lists can be before they are too difficult to evaluate. Another piece of future work is to give the same randomly selected items to multiple people

[7] Actually, this η could be different depending on the opponent being c, d, or e (assuming topics are random). There is thus a mixture, but the marginal probabilities will be the same when integrating out the other processes.

[8] Although we made the values for $p_2(a, b)$ more exact using a Monte Carlo method.

[9] There is measured disagreement in the two sub-tests only when comparing CHI and BNS: CHI wins in head-to-head contests against BNS, but BNS is selected over L1LR, D-TFIDF, and COOC more often than CHI is. However, the degree of discrepancy over all is not enough to rise to a level of statistical significance, and so the matter of interpretation is moot.

Sch. a	Sch. b	% a	n	p_1	%a−%b	n_a, n_b	p_2	$-\log B$	p-value
L1LR	COOC	70	23	0.115	9	69,76	0.300	3.4	0.151
L1LR	DTF	60	22	0.503	31	70,77	0.000	8.3	**0.002**
L1LR	CHI	95	26	0.000	46	66,81	0.000	17.1	**0.000**
L1LR	BNS	94	21	0.000	50	71,78	0.000	15.8	**0.000**
COOC	DTF	62	25	0.383	11	74,74	0.257	2.3	0.327
COOC	CHI	95	27	0.000	26	72,80	0.001	17.0	**0.000**
COOC	BNS	75	24	0.077	43	75,75	0.000	10.2	**0.000**
DTF	CHI	67	26	0.302	25	73,81	0.001	8.1	**0.003**
DTF	BNS	79	26	0.057	26	73,73	0.001	9.8	**0.001**
CHI	BNS	80	28	0.109	-2	79,71	0.785	2.5	0.297

Table 8: Comparing Scheme Performance, Head-to-Head. By column: the name of scheme a; the name of scheme b; rate at which a beat b in the number n of head-to-head a v. b contests; associated $p_1(a, b)$ value; the rate a beat any third scheme minus that of b's; total number of such third-party contests for scheme a, b; associated $p_2(a, b)$ value; negative logarithmic transform of $B(a, b)$ score, and $B(a, b)$'s p-value. Scheme victories scored by counting the number of times they were picked first. Direct contests between a and b with no first-place winner are dropped. Contests between either a or b vs. any of the others are counted as win if a or b was picked first and a loss otherwise. Contests where the content validation step (on the paragraphs) failed (as in paragraphs deemed not about the target topic) also dropped; thus the ns (and power) vary by test. p-values suggesting scheme a significantly outperforms scheme b highlighted.

to assess inter-rater reliability, i.e., whether there is much noise introduced by readers' assessments. We want this to be low, since that strengthens the argument that the results (e.g., that a given process generates better word lists) are generally true rather than individual-specific. Sample size is also, of course, of concern. To bring greater power to these tests, and to convince ourselves of the generalizability of our results, we plan on running a wider array of human validation studies with larger number of participants to better substantiate our preliminary findings.

5. CONCLUSIONS

Social scientists engaged in media studies can find many useful tools in the literature of text classification techniques. By allowing quick summarization of words and concepts as they are portrayed in the media, an approach from machine learning can provide a useful starting point on the analysis of how news media may shape readers' perceptions of the world. But we need to assess quality of these instruments by means other than simple classification performance. As an alternative, we conducted human evaluations of word lists. We find it is important to avoid evaluating feature selection schemes solely on standard predictive error metrics when the objects of interest are the features themselves.

Over these news data, three selection schemes we have studied perform more or less equivalently superior to the remaining two in terms of predictive performance: sparse logistic regression (L1LR), Delta-TF-IDF and the χ^2 likelihood ratio score. However, when the list quality of these three schemes is assessed by human readers that we find the L1LR scheme pulls distinctly ahead, coming out on top in both dimensions of qualtiy. The added computational intensity is likely worth the improved quality, though this claim requires further investigation.

Additional further research will concentrate on solidifying our human evaluation surveys with larger sample sizes; addressing new and more informative schemes building off of regularized classifier training; speeding computation; and seeking greater input from true intended end users to ensure that design priorities maintain productive alignment. It is hoped that our approach can guide the development of machine learning algorithms and performance metrics that are well attuned to human requirements in word imaging tasks for media studies.

Acknowledgments. We would like to thank Jasjeet Sekhon, Saheli Datta, and Vivian Viallon for useful comments and suggestions for improving this paper. We also thank the undergraduate students involved in our human evaluation: Isabelle Chen, Olivia Kim, Alec Kennedy, Ziang Xi and Alli Xi Zheng. Support from the National Science Foundation, CDI grant SES-083553, is gratefully acknowledged.

6. REFERENCES

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[2] B. Berelson. *Content Analysis in Communications Research*. American Book-Stratford Press, 1952.

[3] R. M. Entman. *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. University of Chicago Press, 2004.

[4] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.

[5] A. Genkin, D. Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[6] M. Gentzkow and J. M. Shapiro. What drives media slant? Evidence from U.S. daily newspapers. Working Paper 12707, National Bureau of Economic Research, November 2006.

[7] B. Gunter. The quantitative research process. In K. B.

Jensen, editor, *A Handbook of Media and Communication Research*, pages 209–234. Routledge, 2002.

[8] R. Jervis. *Perception and Misperception in International Politics*. Princeton University Press, 1976.

[9] K. Koh, S.-J. Kim, and S. Boyd. An interior point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

[10] L. Lee and S. Chen. New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents. *Lecture Notes in Computer Science*, 4031:1280, 2006.

[11] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[12] J. Martineau and T. Finin. Delta tfidf: An improved feature space for sentiment analysis. In *Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media*, San Jose, CA, May 2009. AAAI Press. (poster paper).

[13] F. Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005.

[14] R. North. *Content Analysis; A Handbook with Applications for the Study of International Crises*. Northwestern University Press, 1963.

[15] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135, 2008.

[16] Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024, 2005.

[17] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[18] J. L. Solka. Text data mining: theory and methods. *Statistical Surveys*, 2:94–112, 2008.

[19] A. Thompson. The news media and international relations: Experience and the media reality. *Canadian Journal of Communication*, 13(1):53–54, 1988.