No (Privacy) News is Good News: An Analysis of New York Times and Guardian Privacy News from 2010–2016

Karthik Sheshadri
Department of Computer Science
North Carolina State University
Email: kshesha@ncsu.edu

Nirav Ajmeri
Department of Computer Science
North Carolina State University
Email: najmeri@ncsu.edu

Jessica Staddon
Department of Computer Science
North Carolina State University
Email: jessica.staddon@ncsu.edu

***Abstract*—Privacy news influences end-user attitudes and behaviors as well as product and policy development, and so is an important data source for understanding privacy perceptions. We provide a large-scale text mining of privacy news, focusing on patterns in sentiment and keywords. This is a challenging task given the lack of a privacy news repository and a ground truth for sentiment. Using high-precision data sets from two popular news sources in the U. S. and U. K., the New York Times and the Guardian, we find negative privacy news is far more common than positive. In addition, in the NYT, privacy news is more prominently reported than many world events involving significant human suffering. Our analysis provides a rich snapshot of this driver of privacy perceptions and demonstrates that news facilitates the systematization of privacy knowledge.**

I. Introduction

News has been shown to impact user perceptions and behavior. The sentiment of communication can influence behavior (e.g., "fear appeals", [30]), increase user concern (e.g., negative press about the Facebook News Feed launch [26]), and influence perceptions of public opinion [38], [23]. The volume and prominence of news has also been found to be associated with attitudinal changes (e.g., [52]). The end-user impact is likely substantial in the specific case of *privacy* news because news publications are a prevalent source of privacy information. Major news publishers cover privacy-related incidents and regulatory and legal events, review privacy-related products and provide instructional guides for users seeking to protect their privacy.

News also appears to influence policy and product development. Regulators and legislators often cite news as justification for their positions (e.g., [5], footnote 52) and negative press around products is often followed by significant product changes (e.g., the end of the Quora "Views" feature, [2]).

Given the influence of news on end-users, policy and technology development, it is useful to study trends in privacy news. We provide a large-scale text mining of privacy news using two influential news sources, the New York Times (NYT)[1] and the Guardian[2]. In addition to the large volume of privacy news made available by these two publications (Section IV), our choice is motivated by their well documented influence on public attitudes and perception [6], [31], [20], [13]. Motivated by the aforementioned evidence that sentiment (i.e., the attitude or feeling of articles), content (e.g., keywords) and presentation (e.g., page location) of news articles impact user attitudes and behaviors, we focus on measuring these attributes.

This analysis presents three challenges that are specific to the privacy context. First, the sentiment of an article may reflect both objective factors (e.g., harms such as financial loss) as well as subjective factors (i.e., the framing and presentation of the article). This is not specific to the sentiment of privacy news, and for tasks such as predicting movie review ratings from review content, it is not important to distinguish these sources of sentiment. However, we argue that the subjective aspects of sentiment are very important to the privacy context as they relate to aspects of privacy harm that have proven hard to quantify or even define (e.g., [11]). For example, while many have expressed surprise when the aggregation of public information leads to a privacy outcry (e.g., [26], [57]), the sentiment around news and user feedback regarding such events indicates a harm of some kind. We provide evidence of subjective sentiment by *calibrating* privacy news sentiment against the sentiment of random selections of news and non-privacy events and topics.

A second challenge is the absence of a verified ground-truth for privacy news sentiment. While in many sentiment analysis tasks a ground-truth source is available (e.g., movie ratings for reviews [43], stock price changes for company news [32], and congressional votes for political statements [58]) there is no natural ground-truth for privacy news. Instead, we employ human review of sentiment for a sample of articles, multiple independently-trained sentiment classifiers[3], and an analysis of the frequency of sentiment terms from a well-known affect lexicon, to gain confidence in our sentiment measurements.

Finally, a corpus of privacy news is itself challenging to assemble as simply querying news for "privacy" results in many false positives (e.g., "Family of man accused in SC church shooting asks for privacy", [7]). To verify a low false positive rate in our New York Times and Guardian data sets, we employ two independent coders to review a sample of articles for privacy relevance, finding high precision in both. Given that news publications have a strong incentive to broadly cover events, we expect the sets have high recall as well.

Using a high-precision data set of 682 New York Times (NYT) articles from 2010–2016 we find privacy news is predominantly negative according to a majority aggregate of classifiers (Figure 1) and significantly more negative than randomly selected news and events involving significant human suffering like the Zika virus[4] (Table IV). In addition, we find evidence that negative privacy articles are more prominently reported in that they are more likely to be front page news than both randomly selected negative articles and a selection of events of global concern (Figure 7).

[1] A U.S.-Based publication, www.nytimes.com

[2] A U.K.-based publication, https://www.theguardian.com/

[3] All the classifiers we use are trained on separate sets of movie reviews; more in Section III.

[4] The pun in the title refers to this finding; privacy news tends to not be positive, or "good".

We also measure the sentiment of $1,000$ Guardian articles and find that while four independently-trained classifiers agree that negative privacy news is more common than positive privacy news, two of the classifiers find neutral privacy news to be most common. The negativity is evident at the unigram level in that we find negative sentiment words are more common in Guardian privacy news than in randomly selected articles.

Taken together, these NYT and Guardian findings are evidence of the subjective sentiment often call *framing* bias, meaning a tendency toward negative story framing, and *selection* bias, meaning negative privacy news is more prominently reported than a wide array of other negative news. When combined with the previously mentioned research on the impact negative and prominent news has on human perceptions, this suggests news contributes strongly to privacy concerns and the stature of privacy as an end user issue.

We also provide a broad analysis of tags and keywords associated with privacy news and other categories. This analysis allows the tracking of entities (e.g. companies) mentioned in coverage of privacy events over the time period, as well as keywords that are indicative of privacy. We find that these keywords also reflect most of the top user concerns reported through large-scale surveys of user privacy perceptions. Hence, our analysis shows that news facilitates the systematization of knowledge of privacy perceptions and events.

We emphasize that our goal is not to determine whether the measurements found, or potential user impact, are or are not merited. Rather, the evidence that news sentiment, keyword patterns and prominence influence user attitudes and behavior, motivates the task of *measuring and describing privacy news*; the goal of this paper. In summary, our major contributions are:

1) A descriptive analysis of sentiment, entities and keywords associated with more than six years of privacy news reporting.
2) Evidence of framing and selection bias in privacy news that shows privacy events are treated comparably to issues of global concern by the New York Times and the Guardian.
3) Evidence that news facilitates the systematization of privacy knowledge by identifying privacy concerns and events.

II. Related Work

We mine the text of privacy news to understand sentiment and discriminative keywords. Our work is closest to [40], which uses articles from four U.S. news publishers to examine the frequency of privacy-related articles about various types of technology in the years 1985–2003. The data set in [40] is gathered by keyword search ("privacy" and technology keywords such as "cookies" and "rfid") and precision of the set is not verified. Our analysis builds on [40] in that we use a recent, high-precision data set and focus on trends in sentiment, discriminative keywords and the prominence of privacy articles, in addition to term frequencies.

The focus of our analysis is motivated by previous research on the attitudinal and behavioral impact of news and privacy information. For example, [26] finds that user privacy concerns increases in reaction to negative media coverage of the Facebook news feed launch and negative privacy/security-related messages, also known as "fear appeals", are found to increase privacy-related behavioral intent in in the corporate context [30]. In addition, news in general has been found to influence perceptions of public opinion [38], [23] and the volume and prominence of news has also been found to be associated with attitudinal changes (e.g., [52]). Reader interest in negative news is found in [59]. These works all motivate our focus on sentiment, prominence, and discriminative keywords, as their findings suggest these attributes influence attitudes and behavior.

Our work is also motivated by efforts to gather and analyze privacy incidents (e.g., [16], [47]) in that we show how semi-automated analysis privacy news supports the characterization of privacy incidents over time.

There are also general results in news research (i.e. not focused on privacy) that are relevant to ours. In particular, news has been found to be biased (e.g., [48]) and negative in many contexts, particularly when on the front page [55]. We extend this work by showing a negative trend in the case of privacy news.

Similar to ours, [19] develops a sentiment classifier and applies it to a general sample of news, finding that both positive and negative sentiment are more common than neutral. Our findings for privacy news differ in that negative sentiment is more common that the positive.

Finally, the natural language processing and machine learning tools we employ have been applied in many other areas.For example, sentiment classification has also been applied to social media, including Twitter (e.g., [39]) and Facebook (e.g., [41]). In addition, text mining is increasingly applied to privacy. It has been used to help ensure privacy policies are enforced correctly (e.g., [10]), to sanitize text (e.g., [29], [56]), to identify patterns in app permission requests (e.g., [15]) and to understand content sensitivity [45].

III. Background and Approach

As mentioned earlier, our primary goal is a descriptive analysis of privacy news. Motivated by evidence that attitudes and behaviors are influenced both by sentiment (e.g., [30]) and the volume and prominence of news (e.g. [52]), we focus on understanding the sentiment, keyword patterns and placement of privacy news. We also study keyword patterns to test our hypothesis that news analysis supports ongoing efforts to automate the collection and analysis of privacy incidents (e.g., [42], [16], [47]). Our approach makes use of well-established text mining tools that we describe in the following.

Sentiment analysis. We classify sentiment by considering it as a topic-based classification problem and using standard text classification algorithms (e.g., as described in [33]). We rely most on the Stanford Core NLP parser [53], [34] (widely considered the state of the art in sentiment analysis), however, in many cases we also use the well-established tools, the Python NLTK sentiment analysis API[5] and Matlab implementations of sentiment classifiers based on Naive Bayes and Support Vector Machines (SVM) for additional measurements.

Our evaluation is based on a division of articles into three classes: positive, neutral, and negative. We reviewed samples of all classifier output and found high agreement between human-assessed sentiment and classifier output (more on this in Section V).

The Stanford sentiment parser [53] relies on a manually labeled 'treebank' of approximately 12k English language sentences extracted from movie reviews (of comparable length to news articles). Each sentence is decomposed into its constituent parts of speech, and the underlying keywords are labeled with a sentiment score. A recursive neural network is trained on this corpus to map each input document into a sentiment range of 0 to 4. We utilize the open source Java implementation of CoreNLP, which provides a discrete sentiment rating on the scale described above, and then we use thresholds to separate articles into three classes: $0--1$ negative, 2 neutral, $3--4$ positive.

We also implement standard Naive Bayes and Support Vector Machine (SVM) approaches to sentiment analysis in MATLAB. Our

[5]http://www.nltk.org/api/nltk.sentiment.html

classifiers are trained on the Cornell polarity data set [44] which consists of $2,000$ movie reviews (again, of comparable length to news articles), each labeled on a scale of -4 to 4. Again, we manually separate our data into three classes using the same scale as for CoreNLP: -4 to -0.81 negative, -0.8 to 0.8 neutral, and 0.81 to 4 positive.

To determine the sentiment class of articles, we first preprocess the data via stemming and the removal of stop words. We then extract bigrams from each article, constructing a feature vector which is used to calculate the standard posterior probability in the case of Naive Bayes. The class confidence values in the case of SVM are calculated as: $d_{FV} = \sum_{i=1}^{m} \alpha_i SV_i \cdot FV + b$, where d_{FV} is the distance of feature vector FV to the separating hyperplane defined by the m support vectors SV, $\{\alpha_i\}_i$ are the learned weights of the SVM, and b is the bias (as in [33]).

SIGNIFICANCE TESTING AND EFFECT SIZE. Our data consistently fail the standard one sample Kolmogorov-Smirnov normality test [35] at the 0.05 level and so are unlikely to be normally distributed. Hence, we use nonparametric significance testing, specifically, the two sample Kolmogorov-Smirnov test [35] to test whether two sample groups belong to the same underlying probability distribution function. We use the posterior probability of negative class membership as our distribution, and make the probability space into ten discrete intervals of equal length. This yields a normalized probability space within which sample group pairs may be compared using the Kolmogorov-Smirnov test.

With larger data sets, statistical significance is more likely and it is important to consider the size of the effect in addition to significance. The more well-established tools for measuring effect size (e.g., Cohen's d) are best used with normally distributed data, so we emphasize visual representations of our data to demonstrate practical significance and employ a nonparametric effect size measure, Cliff's delta (e.g., [36]) when visualizations are difficult to provide.

BIAS MEASUREMENT. We term the trend toward negativity in news, *framing bias*. To test whether framing bias exists in privacy news we compare the sentiment of privacy news (using classification techniques described in this section) and the sentiment of various samples of non-privacy news, and we evaluate whether sentiment words in an established affect lexicon[6]are more closely associated with privacy news than other news categories.

To help calibrate the degree of framing bias, that is, the degree of negativity of privacy news, we also measure *selection bias*, meaning, the tendency to report negative privacy events more prominently than other negative events. We measure prominence by a news article's page number and compare sentiment categories of privacy news with those of randomly selected news and non-privacy news.

KEYWORD DETECTION. The problem of keyword detection in general text mining contexts has been well studied. Approaches range from simple histogram vectorization [50], to Pivoted-Normalized document length [51] and query weighting schemes [8].

We make use of Term Frequency-Inverse Document Frequency (tf-idf) [46], which measures how important a word is to a document based on frequency of occurrence in the document relative to frequency in the whole corpus. Keywords are detected by sorting all terms in a privacy news corpus according to their tf-idf, and reporting the top k of those alongside the top k from that of a random news corpus.

[6]Available at: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html\#lexicon

Classifiers can use utility measures like tf-idf to decide classes. In our context, the tf-idf of article terms can be used to determine whether an article is an example of privacy news. More generally, given an unobserved dependent variable y, discriminative classifiers model the dependence of y on an observed variable x. This is done by modeling the posterior $P(y|x)$ directly, or by learning a mapping from x to y.

Information gain [24] is another utility measure. Each news article can be represented as a vector in an m-space, in which every distinct n-gram represents a dimension.

We define the utility of n-gram, x_i, to be:

$$IG(T, x_i) = H(T) - \sum_{i=1}^{m} \frac{|A \in T | x_i \in A|}{|T|} H(\{A \in T | x_i \in A\}) \quad (1)$$

where $T = \{A\}$ denotes the set of training example articles, H the information entropy of T and there are m, n-grams.

In this paper, we use the random forest [9] algorithm, in conjunction with both approaches to measuring term utility (tf-idf and information gain) to learn n decision trees, each trained in a greedy fashion on a randomly chosen subset of the data. Each split node of a decision tree chooses the n-gram that maximises utility among a randomly chosen subset of the n-gram set.

IV. DATA SETS

We use news articles gathered from The NYT Developers' API [3] and The Guardian API [22]. The NYT and the Guardian were chosen primarily for their popularity in the US (the NYT) and the U.K. (the Guardian), and the existence of APIs for both sources. The articles fall into three distinct categories: privacy news, articles about events of global concern that are not privacy-related (often referred to as the "non-privacy" data set) and randomly selected sets of news articles. We describe the NYT data sets in Section IV-A, the Guardian data sets in Section IV-B and the precision of each privacy news set in Section IV-C. The non-privacy set is gathered from the NYT and described in Section V-B.

A. NYT Data Sets

We queried the NYT Developers' API (Article Search API v2) for the search word 'privacy' for dates from January 1, 2010 to September 30, 2015. The API looks for a search word in the article headline and body text and returns a short summary of each article and the lead paragraph, we base our analysis on the summaries and lead paragraphs.

Our initial API call returned 13,077 articles, we then applied three filters to remove non-privacy articles: (1) we restrict the source to "The New York Times", removing sources such as Reuters and Associated Press, (2) to avoid author bias, we restrict our data set to articles whose "Type of material" is "News", removing blogs, editorials, and op-eds, etc., and (3) we restrict our set to the articles that are tagged with the subject "privacy" (tags include subject, organizations, persons and geographical locations).

These filters reduced our data set to 682 privacy articles. In our analysis we compare against disjoint, randomly selected sets of NYT news articles of 500 articles each, denoted Random1 and Random2 in the following.

We also used the API to gather articles on the following non-privacy topics: racism, China's economy, China's One Child Policy, a January 2016 snow storm in the eastern United States, the Syrian refugee crisis and the Zika virus. We describe these data sets in more detail in Section V-B.

B. Guardian Data Sets

The Guardian API provides access to a total of $\approx 22K$ articles tagged 'privacy', dating back to April 1986. We extracted the URLs of a random sample of 400 articles from 2010 to 2016 for which we evaluated the precision as described in Section IV-C. In Sections V and VI we analyze the sentiment and keywords of a set of $1,000$ privacy and $1,000$ randomly selected articles, using the complete article text. The Guardian API categorizes articles in a more granular manner and we could not find a category or set of categories comparable to "news" in the NYT data set, hence we chose to not filter articles into news/non-news classes based on article tags.

C. Data Set Accuracy

To gain confidence that the articles in the NYT and Guardian data sets are *privacy* news articles, we manually reviewed samples from each set. In particular, to estimate the precision we coded random samples of 225 NYT articles and 200 Guardian articles. Each sample was coded by two people; one person coded both sets. We employ a simple coding scheme; articles are classified as to whether or not they are news and whether or not they are about privacy, using the following guidelines:

News articles: Articles that cover recent events in such a way that they grow the public understanding of the events (i.e. reporting on the events rather than just citing them).

Privacy Articles: Articles about privacy in the context of humans and digital data, for which privacy is core to the piece and is likely to draw readers to the story.

For both data sets, the coders first coded 100 randomly selected articles independently. They then met to discuss disagreements and refine their understanding of the guidelines, and then coded the remaining articles (100 in the case of the Guardian and 125 for the NYT). After the second round of coding, the coders again met to discuss disagreements, resolving many of them. The final inter-coder agreement, as measured by Cohen's κ [61], is $\kappa = 0.92$ on the 200 Guardian articles and 0.95 on the 225 NYT article sample. The simple coding scheme likely contributed to the high inter-coder agreement.

The two coders both measured news precision of 0.685, and privacy precision of 0.97 and 0.965, respectively, on the 200 Guardian articles. They measured news precision of 0.9 and 0.898, respectively, and privacy precision of 0.96 and 0.929, respectively, on the 225 NYT articles. The coder disagreements were almost equally split between the news classification task (5 disagreements) and the privacy classification task (7 disagreements).

The primary coder of the Guardian articles coded a subset of 400, finding news precision of 0.6425 and privacy precision of 0.9625.

While we replicate the NYT article analysis with the Guardian data set in most cases, the NYT data set is analyzed more thoroughly overall both because of the expressive tagging supported by the NYT API and because it is the data set with the highest precision in terms of both privacy and news.

V. News Sentiment

Table I summarizes our sentiment measurements for privacy news and randomly selected articles. Across two publishers we find evidence that privacy news is significantly more negative than randomly selected news.

For the NYT data set in particular, the negative trend in privacy versus random news is very pronounced; the proportion of negative privacy news exceeds the proportion of negative randomly selected news by almost 40% (Table I). We find almost uniform agreement across our four independent classifiers for this data set, across two disjoint random samples. It is a reasonable conclusion from these data that on average an NYT privacy news article is *expected* to have negative sentiment.

While the Guardian privacy dataset is predominantly neutral, the percentage of negative sentiment privacy articles is still significantly above the corresponding fraction of the randomly selected articles (at least 7% according to any classifier).

In Table II we explore the conjecture that news articles are less negative than other categories of articles. We find this is only the case for the OpEd articles in NYT sample, which are significantly more negative than privacy news.

Human-Review of Sentiment. To gain confidence in the classifier output, one of the authors reviewed samples of 100 NYT and 100 Guardian privacy samples, and classified each article as either negative or not negative in sentiment. The samples were balanced in terms of (classifier-determined) sentiment; in each set, 50 were determined to have negative sentiment by a majority of classifiers (as described in Section III) and 50 did not. The human reviewer did not have the sentiment measurements when reviewing. The human review largely agrees with the classifier measurements. In particular, on the NYT sample, when compared with the human reviewer's classifications, our classifiers achieve an accuracy of 0.78, with (precision, recall) pairs of $(0.70, 0.775)$ on the negative articles and $(0.84, 0.78)$ on the other articles. Similarly, with respect to the Guardian sample, the classifiers achieve an accuracy of 0.72, with (precision, recall) pairs of $(0.67, 0.73)$ on the negative articles and $(0.77, 0.72)$ on the other articles. This level of accuracy is comparable to that achieved in other sentiment analysis tasks such as predicting movie ratings [43] and stock market reaction to company news [32].

Examples of articles determined to be negative or not negative by the human reviewer and a majority of classifiers are in Table III. For the negative articles, the table also includes words and phrases that were inputs both to the classifier and human review and that the human reviewer considered evidence of negative sentiment.

Agreement between Classifiers. We also measured agreement between our classifiers to gauge the consistency of our results. We treat each classifier as a rater in the spirit of Cohen [61], and calculate pairwise Kappas between our classifiers. The median Kappa is 0.56, which is considered "good agreement" [14]. The corresponding Kappa between a majority aggregate of our classifiers and the human coder described above is 0.58.

A. Sentiment by entity

The articles in the NYT privacy data set are tagged by the NYT with organizations, people, and geographical locations. We filter based on these tags, and describe the trends below.

Organizations. 173 of the 682 news articles in the NYT privacy data set are tagged with one or more organizations. Overall, 142 unique organizations are tagged, of which 32 organizations appear more than once, with Google having the most support, followed by Facebook and the Federal Trade Commission. Figure 2 shows a line plot of year-wise number of articles for some of the prominent organizations. Figure 3 shows a word cloud of support for organizations.

Of the 142 organizations, 37 appear in positive articles at least once, with Google appearing most, followed by Facebook. 93 organizations appear in negative articles at least once, and 22 of these 93 organizations appear in negative articles more than once. Google is the organization most frequently tagged in negative news articles. It

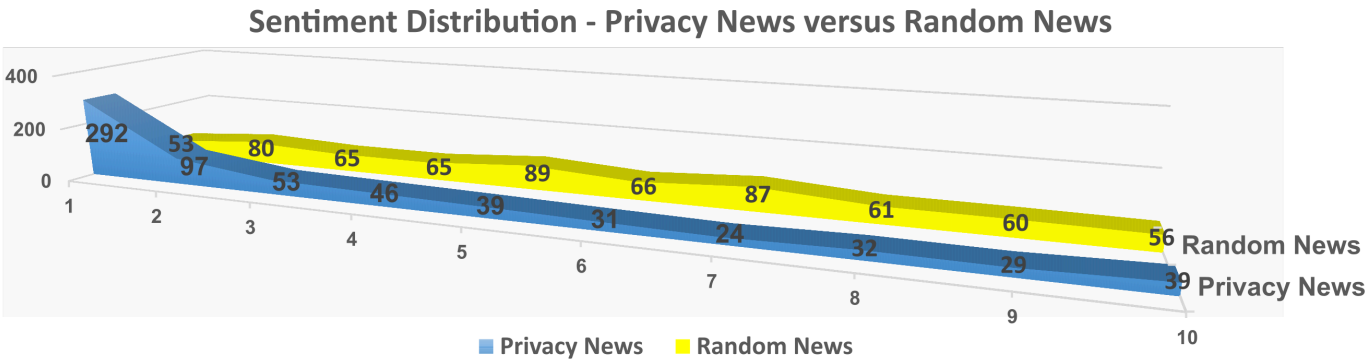

Fig. 1: Sentiment distribution of the New York Times privacy news dataset (682 articles) in comparison to 682 randomly selected New York Times news articles. Measured sentiment increases in positivity as scores increase. Random news articles are almost uniformly distributed across the range of sentiment scores, but privacy news articles are heavily biased towards the negative end of the spectrum.

TABLE I: A quantitative comparison of privacy vs randomly selected news on our NYT and Guardian data sets. The percentage of negative articles is often higher for privacy news samples than for random ones and is greater than the percentage of positive privacy news according to all classifiers and for both news sources.

Publication	Data Sample	Classifier	Negative	Neutral	Positive	Data Set Size
NYT	NYT Privacy	Stanford	470 (68.91%)	122 (17.89%)	90 (13.20%)	682
		NLTK	446 (65.39%)	152 (22.28%)	84 (12.31%)	
		Naive Bayes	386 (56.60%)	144 (21.10%)	152 (22.30%)	
		SVM	381 (55.86%)	140 (20.51%)	161 (23.60%)	
	NYT Random1	Stanford	167 (33.40%)	215 (43.00%)	118 (23.60%)	500
		NLTK	163 (32.6%)	201 (40.20%)	136 (27.20%)	
		Naive Bayes	156 (22.90%)	399 (58.50%)	127 (18.60%)	
		SVM	167 (24.49%)	393 (57.60%)	122 (17.88%)	
	NYT Random2	Stanford	205 (41.00%)	190 (38.00%)	105 (21.00%)	500
		NLTK	186 (37.20%)	195 (39.00%)	119 (23.80%)	
		Naive Bayes	225 (45.00%)	199 (39.80%)	76 (15.20%)	
		SVM	201 (50.24%)	200 (50.00%)	99 (24.76%)	
Guardian	Guardian Privacy	Stanford	331 (33.00%)	642 (64.20%)	27 (2.70%)	1,000
		NLTK	347 (34.70%)	611 (61.00%)	42 (4.20%)	
		Naive Bayes	297 (29.70%)	633 (63.30%)	70 (7.00%)	
		SVM	321 (32.10%)	563 (56.30%)	116 (11.60%)	
	Guardian Random	Stanford	275 (27.50%)	673 (67.30%)	52 (5.20%)	1,000
		NLTK	232 (23.20%)	684 (68.40%)	84 (8.40%)	
		Naive Bayes	226 (22.60%)	576 (57.60%)	198 (19.80%)	
		SVM	267 (26.70%)	563 (56.30%)	170 (17.00%)	

TABLE II: A comparison of sentiment (as measured by a majority aggregate of our analyzers) between the NYT privacy articles that are tagged by the NYT as News, OpEd, Blogs and Editorials. In the privacy blog and editorials data set, we measure less negative sentiment than in the privacy news set.

Type	Negative %	Neutral %	Positive %	# Articles
OpEd	80.88	11.76	7.35	68
News	68.91	17.89	13.20	682
Blog	50.12	35.91	13.96	401
Editorials	45.31	48.44	6.25	64

is followed by the National Security Agency (NSA) and Facebook. Figure 3 shows a word cloud of support for organizations.

People. 76 articles of 682 news articles in the NYT privacy data set are tagged with one or more individuals. Overall, 101 unique people are tagged, with 9 appearing more than once. 67 people of the 101 appear in negative articles at least once, and 3 appear in more than one negative article. "Snowden Edward J" is the most common person tag and is associated with the largest number of negative news articles.

Geographical Locations. 215 articles of 682 news articles in the NYT privacy data set are tagged with geographical locations. Overall, 95 unique locations are tagged, of which 35 locations appear more than once. 83 of the 95 locations appear in negative articles at least once, and 33 locations appear in negative articles more than once. "Europe" is the most common location both overall and in the negative news articles, followed by the United States and Germany.

TABLE III: Examples for which the sentiment classifiers and human reviewer agreed. Article excerpts are taken from the text available to both the classifiers and the reviewer.

Title	Classification	Negative Words or Phrases	Source and Date
F.B.I. Violated Rules in Obtaining Phone Records, Report Says	Negative	"...The FBI improperly obtained..."	NYT, 1/20/2010 [49]
Critics Say Google Invades Privacy With New Service	Negative	"...firestorm of criticism"	NYT, 2/12/2010 [25]
Samsung's voice-recording smart TVs breach privacy law, campaigners claim	Negative	"complaint", "unfair", "deceptive"	Guardian, 2/27/2015 [18]
Facebook's privacy policy breaches European law, report finds	Negative	"violation"	Guardian, 2/23/2015 [17]
A Call for a Federal Office to Guide Online Privacy	Not Negative	NA	NYT, 12/16/2010 [60]
Another Try by Google to Take On Facebook	Not Negative	NA	NYT, 6/28/2011 [37]
David Bowie's family thank fans for tributes and renewrequest for privacy	Not Negative	NA	Guardian, 2/14/2016 [12]
Public to get online access to US workplaces' injury and illness records	Not Negative	NA	Guardian, 5/11/2016 [21]

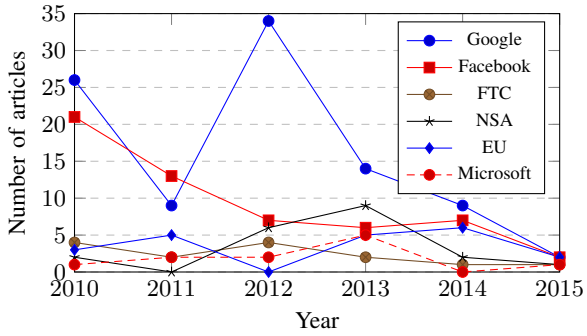

Fig. 2: Year-wise number of articles for the most commonly mentioned organizations in the NYT data set. *Only until September 2015.

B. Sentiment by Topic

We used the NYT developers' API, and a manually generated set of search terms[7] to recover articles for particular news topics. These events were chosen because they involve significant human suffering and hence are widely regarded as highly negative in nature, have strong support in our news sources and are not particularly related to privacy. Additionally, publishing in these topics has been shown to impact public opinion and behavior as a result of their sentiment [23]. Many of these events made "year-in-review" lists from other publishers (e.g., [4]). Since none of these topics is privacy-focused, we refer to the group as a whole as being the "non-privacy" data set in the following.

Some of these topics evolve over several years and article sentiment may change over that time (e.g., China's One Child Policy, which, when it ended, resulted in articles with less negative sentiment). To gauge the initial sentiment around the topics, we limit sentiment analysis to approximately the first month of the topic. The exception to this is the topic 'racism' which does not have a clear 'first month' (much like the topic, 'privacy') and so we considered all articles from 1/1/2010–12/15/2015.

We summarize findings by topic below. As shown in Table IV privacy news is more negative than most other categories (as computed by a majority aggregate of classifiers), with differences that are significant and have medium/large effect sizes between the privacy sample and random samples. The two topics that are more negative than privacy, racism and the Syrian refugee crisis, have negligible and small effect sizes.

China's Economy. China's economy recently witnessed a considerable downturn. We queried the NYT with the search term "China" and "economy" for the dates January 1, 2016 to February 8, 2016, and identified 716 articles. 65.22% of these articles are negative in sentiment.

China's One-Child Policy. The NYT developers' API returned 40 articles for the search term "one child policy" for the dates January 1 to February 1, 1978. 65% of these articles are negative in sentiment. This collection includes articles about the end of the policy; such articles are expected to be more positive in sentiment.

Paris Terrorist Attacks. On November 13, 2015, coordinated attacks were carried out in Paris by militants [8]. We queried the NYT for the terms "Paris" and "attack" for the period November 13 to December 13, 2015 and identified 929 articles. 66.23% articles of these articles are negative in sentiment.

Racism. Racism is a continuing issue of global concern. The NYT developer's API returned 3,266 articles for this search term during the dates January 1, 2010 to December 15, 2015. 73.67% of these articles are negative in sentiment.

2016 East Coast Snowstorms. The eastern United States was hit by a category-5 snow storm [9] in January 2016. An NYT search for the query "snow storm" returned 232 articles for the dates from January 1, 2016 to February 8, 2016, of which 53.45% are negative in sentiment.

Syrian Refugee Crisis. At the time of writing, a civil war that began in 2011 is driving many Syrian residents to flee the country and seek refuge elsewhere. We searched the NYT for "syria refugee" during the period June 1 to July 1, 2011, and identified 17 articles. 70.5% of these articles are negative in sentiment.

Zika Virus. We searched the NYT for articles on Zika virus for dates January 1, 2016 to February 8, 2016, and identified 409 articles. 60.64% of these articles are negative in sentiment.

[7]While we did not thoroughly evaluate topic recall, we used a few round of keyword iteration to gain confidence in coverage.

[8]https://en.wikipedia.org/wiki/November_2015_Paris_attacks

[9]https://en.wikipedia.org/wiki/January_2016_United_States_blizzard

TABLE IV: A comparison between the sentiment of privacy news and of articles related to racism, China's economy and One Child Policy, a recent snow storm in the USA, the 2015 Paris terrorist attacks, the Syrian refugee crisis, the Zika Virus, and two random sets. All data sets are from the NYT and sentiment is measured with Stanford's Core NLP parser.

Data Set	Negative %	Neutral %	Positive %	Number of Articles	p value	Cliff's delta
Racism	73.67	11.26	15.07	3,266	< 0.05	small
Syrian Refugee Crisis	70.50	17.64	11.76	17	< 0.05	negligible
Privacy	68.91	17.89	13.20	682		
China's economy	65.22	22.21	11.59	716	< 0.05	negligible
Paris Attacks	66.23	24.14	9.63	929	< 0.05	negligible
China's one-child policy	65.00	15.00	20.00	40	< 0.01	small
Zika Virus	60.64	29.83	9.53	409	< 0.01	small
2016 US Snow Storm	53.45	34.05	11.64	232	< 0.01	small
Random1	33.40	43.00	23.60	500	< 0.01	large
Random2	38.00	41.00	21.00	500	< 0.01	large

C. Framing and Selection Bias

Table IV shows the percentage of New York Times privacy news that is negative exceeds both the percentage of negative randomly selected news and the percentage of negative news covering events involving significant human suffering. The difference holds across multiple classifiers and two data sets, and so is evidence of framing bias in privacy news.

We also consider the placement of privacy news within a publication as an aspect of selection bias. Figure 4 shows the plot of page number and percentage of positive, negative, and neutral articles in the NYT privacy data set. We observe that the majority of the privacy articles are published on page 1 of the NYT. This is the case for all sentiment classes of privacy news, with the largest class being negative. While coverage of the events of global concern listed in Section V-B and represented in Figure 6 also tends to be front page news, the cumulative distribution function in Figure 7 shows the tendency is much less pronounced. In particular, about 60% of the negative class of privacy news appears on the first three pages, whereas only 34% of the negative class of non-privacy, and 22% of the negative class of random news appears on the first three pages of NYT.

Figure 5 shows page number and percentage of positive, negative, and neutral articles in the three NYT random data sets combined. Figure 6 shows the plot of page number and count of positive, negative, and neutral articles in the seven NYT non-privacy data sets introduced in Section V-B. We observe that the percentage of negative articles published on page 1 is higher than the percentage of positive articles.

As an additional measure of framing bias, we use an exhaustive list of broadly accepted negative sentiment words[10], and measure the relative frequency of their occurrence in privacy versus random news articles. We use our Guardian data sets; the full text of $1,000$ privacy news and $1,000$ random news articles. While the cumulative occurrence frequency of these negative words in privacy news articles sum to $\approx 91K$ occurrences, the corresponding number for random news is substantially smaller at $\approx 83K$. Cliff's Delta test yields a medium effect size for these distributions.

VI. News Keywords

We present a keyword analysis that characterizes the technologies, organizations and issues in the events and incidents covered in news. The keywords we identify are *discriminative* in that they are closely associated with privacy news and/or a specific category of privacy news. To demonstrate how this analysis supports the systematization of privacy we provide applications to taxonomy development and user perception measurement.

To identify discriminative keywords in the privacy NYT data set, we created three train/test splits of our data by randomly sampling 50% of examples each for our training and evaluation set. Then, we trained random forests using (i) tf-idf utility and (ii) information gain, and measured their accuracy on each test split using 1, 2, 3, and 4 gram based vector spaces.[11]

While we omit the details due to space constraints, information gain consistently outperforms tf-idf in the simple binary classification context and achieves an accuracy of 0.9312 in classifying privacy news articles, hence we rely on the former measure.

TABLE V: The top ($k = 20$) unigram and bigram keywords in the NYT privacy set.

Unigram Keywords	privacy; Facebook; rights; Snowden; rights activist; surveillance; collection; security; release; social; identification; camera; regulator; cookies; Google; track; history; NSA; China; Court; violation
Bigram Keywords	privacy breach; Facebook privacy; privacy policy; rights activist; surveillance camera; browsing history; edward snowden; social security; data collect; location history; street view; identity theft; privacy violation; homeland security; security administration civil liberties; Supreme court; license plate; HTML 5; secure email

Following the process of Section III, we extract the top $k = 20$ keywords from our classifier for the vector cases of $n = 1$ and $n = 2$ and list them in Table V.

Many discriminative keywords returned by our system are unsurprising, for example, the word 'privacy' has a high utility in discriminating privacy articles from non privacy ones. Other keywords reflect import events in privacy news. For example, the term HTML 5 (information gain rank 19) likely surfaces because a Facebook security bug allowed third party app companies such as Farmville creator Zynga to access private user information and then re-distribute it to advertisers and tracking companies [1]. Up to 218 million Zynga users were affected by the bug.

[10] [27], https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html\#lexicon

[11] Preprocessing steps such as stemming and stop word removal are carried out as described in Section III.

Fig. 3: Support for organizations in the complete collection of NYT privacy articles (top), the NYT privacy articles classified as having positive sentiment using the Stanford classifier (middle) and the NYT privacy articles with negative sentiment according to the Stanford classifier (below).

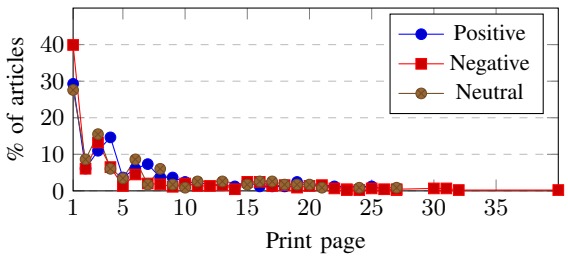

Fig. 4: Sentiment and print page of NYT privacy related articles as measured by a majority aggregate of our analyzers. 38 articles of the 682 privacy related articles in NYT are web-only articles or do not have associated page numbers.

Based on initial experiments with Google News Search, there appears to be a rough correlation between the information gain metric and news volume. (Table VI). For instance, the term 'Facebook privacy' ranked second in terms of information gain in our classifier, and accounted for 62.88% of last year's privacy news, whereas

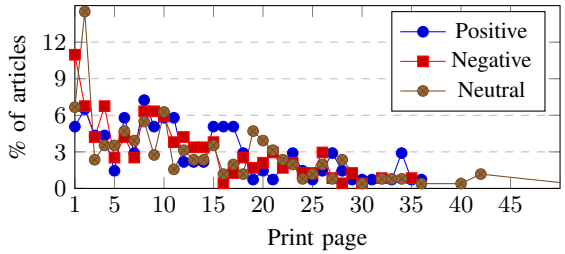

Fig. 5: Sentiment and print page of random NYT news articles as measured by a majority aggregate of our analyzers. A total of 551 of the 1,495 articles are web-only articles or do not have an associated page number.

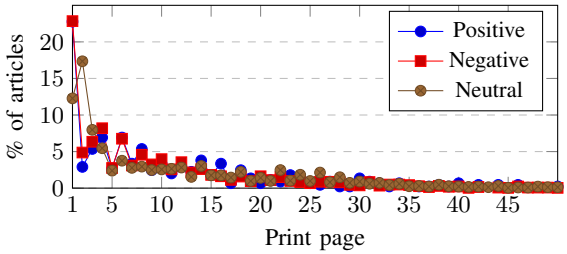

Fig. 6: Sentiment and print page of NYT non-privacy news articles including coverage of a 2016 US snow storm, Apartheid, China's economy, China's one child policy, Paris attacks, Syria refugee crisis, and the Zika virus. Sentiment is measured by a majority aggregate of our analyzers. A total of 4,995 of the 10,438 articles are web-only or do not have an associated page number.

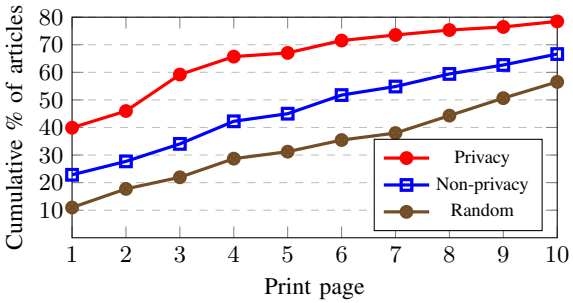

Fig. 7: Cumulative distributions of the New York Times page numbers of the negative classes of privacy news, the non-privacy news (defined in Section V-B) and randomly selected news articles. The negative privacy news is almost twice as likely to appear on the front page as negative non-privacy news, and almost four times as likely as randomly selected news.

'surveillance' ranked fifth (1.28%) and 'civil liberties' ranked 16th (0.04%).

A. Comparison with Other Sources

Because our keyword analysis is based on aggregate data over the years 2010–2015, we compare with sources aggregated over the same time period. In particular, we compare with broad surveys in the U. S. and U. K. as well as a manually gathered list of important privacy events [16].

Our goal in making this comparison is to demonstrate there are compatible *themes* in our data set and the other sources, as opposed

TABLE VI: Information gain rank versus percentage of privacy news articles returned by Google news search for the period 2012-2016. The proportionality pattern demonstrates the impact of our estimated keywords.

Keywords	Information Gain rank	News articles 2012-2016
Facebook privacy	2	62.88%
Surveillance	5	1.28%
Civil Liberties	16	0.04%

to measurable agreement. We do not aim for measurable agreement both because the data are different in form and driving goals, and because, in the case of the surveys, the questions asked of participants vary over the years and by source. That said, because the themes are compatible, we argue that privacy news is a complementary data source for gauging user privacy perceptions.

PRIVACY SURVEYS. Pew Research has reported the results of broad privacy surveys and focus groups in each of 2016, 2015 and 2014[12] Prior to 2014, reported privacy surveys were focused on specific populations and/or specific aspects of privacy, so we do not include them in our comparison. In addition, the European Commision published "Eurobarometer" reports on privacy in 2015 and 2011[13].

Several issues appear in a majority of the reports, in particular, privacy concerns regarding: (1) social media (Pew 2015, 2014, Eurobarometer 2015, 2011), (2) mobile phone providers (Pew 2015, 2014, Eurobarometer 2015, 2011), (3) search engines (Pew 2015, Eurobarometer 2015, 2011), (4) targeted advertising (Pew 2016, 2015, 2014, Eurobarometer 2015, 2011), (5) personal information collection (Pew 2016, 2015, 2014, Eurobarometer 2015, 2011), (6) government (Pew 2016, 2015, 2014, Eurobarometer 2015) and (7) surveillance/monitoring (Pew 2016, 2015, 2014, Eurobarometer 2015). Many of the keywords in Tables V and **??** appear closely related to these issues. For example, for *social media*: Facebook privacy, Buzz, VidMe, Farmville, social, social network; *mobile phone providers*: AT&T, Blackberry; *search engines*: Google, Microsoft; *targeted advertising*: ads, track, cookies; *personal information collection*: collection, browsing history, street view, data collect; *government*: trade commission, TSA, FBI, homeland security, government; and *surveillance/monitoring*: scanners, cameras, x-ray, surveillance.

MANUALLY-CURATED PRIVACY INCIDENTS. In [16], Garfinkel and Theofanos summarize and analyze 42 privacy incidents grouped by Solove category [54]: information collection (10), information processing (10), information dissemination (17), and intrusion[14] (5). Many of the keywords in Tables V and **??** appear to directly represent incidents in [16], for example, "Buzz" and "Street View", and others represent the issues raised by the incidents such as tracking, ads, location history and/or the technology involved, for example, cameras, cookies, apps and facial recognition.

[12]All reports are available at: http://www.pewresearch.org/topics/privacy-and-safety/. Reports reviewed: "Privacy and Information Sharing" (2016), "Americans' Attitudes About Privacy, Security and Surveillance" (2015), "Public Perceptions of Privacy and Security in the Post-Snowden Era" (2014)

[13]All Eurobarometer reports are available at http://ec.europa.eu/COMMFrontOffice/PublicOpinion/. Reports reviewed: "Data Protection" (2015), "Attitudes on Data Protection and Electronic Identity in the European Union" (2011)

[14]Solove's "invasion" category is renamed intrusion in [16]

VII. LIMITATIONS AND OPEN PROBLEMS

Our analysis of trends in sentiment and entities requires a data set of unique articles. To ensure this, we filtered our data sets somewhat conservatively. For example, we filtered the original NYT set of $13,077$ privacy articles (i.e. those tagged with "privacy") available through the NYT API, to those for which NYT is the publication source. This ensures reprints of articles by other publishing sources are removed, but reduces the data set considerably. It is possible that a more sophisticated filter would have resulted in a larger set of unique articles.

It is widely believed that the news in general, and many news publishers in particular, are biased in their reporting. For example, the NYT has admitted to a liberal bias [28]. Given the large readership that both the NYT and the Guardian enjoy, the analysis is useful whether or not the sources are biased. That said, it would certainly be valuable to expand the analysis to more publications to see if the trends continue in other news publications. It would be particularly useful to analyze non-English-language newspapers to see if cultural variation in privacy attitudes found through user studies (e.g. [62]) persist in the media.

VIII. CONCLUSION

We have found patterns of negative sentiment and prominent coverage of privacy news in comparison to several other categories of news included events of global concern, in the New York Times, and similar sentiment results in the Guardian. Given previous research finding that negative privacy accounts impact user attitudes and perceptions of the importance of privacy (e.g., [30], [38]), and the prevalence of news as a privacy information source, our analysis suggests privacy news is an important factor influencing user privacy concerns and perceptions.

In addition, our descriptive analysis demonstrates that the automated analysis of news can inform efforts to efficiently gather and analyze trends in privacy incidents (e.g., [16], [47], [42]).

While our analysis is limited to two publications, the New York Times and the Guardian, we emphasize that both publications are quite influential and have large subscriber bases, making them important to understand as information sources. In addition, through human-review, we have confidence in the precision of our data sources.

REFERENCES

[1] http://www.develop-online.net/news/facebook-id-leak-hits-millions-of-zynga-users/0107956. Accessed: 2016-5-15.

[2] After privacy uproar, quora feeds will no longer show data on what other users have viewed. http://techcrunch.com/2012/08/14/after-privacy-uproar-quora-backpedals-and-will-no-longer-show-data-on-what-other-users-have-viewed/. Accessed: 2016-2-25.

[3] Developer apis. http://developer.nytimes.com/. Accessed: 2016-3-3.

[4] Major global events that shook 2015, 2015.

[5] Opinion 01/2015 on privacy and data protection issues relating to the utilisation of drones. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2015/wp231_en.pdf, June 16, 2015.

[6] S. L. Althaus and D. Tewksbury. Agenda setting and the new news patterns of issue importance among readers of the paper and online versions of the new york times. *Communication Research*, 29(2):180–207, 2002.

[7] Associated Press. Family of man accused in sc church shooting asks for privacy, November 1, 2016.

[8] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *Proceedings of the 34th International ACM SIGIR Conference*, SIGIR '11, 2011.

[9] L. Breiman. Random forests. *Mach. Learn.*

[10] C. A. Brodie, C.-M. Karat, and J. Karat. An empirical study of natural language parsing of privacy policy rules using the sparcle policy workbench. In *Proceedings of the Second Symposium on Usable Privacy and Security*. ACM.

[11] R. Calo. The boundaries of privacy harm. *Indiana Law Journal*, 86(3):1131–1162, 2011.

[12] A. Cowburn. David bowie's family thank fans for tributes and renew request for privacy. *The Guardian*, January 14, 2016.

[13] D. W. Drezner and H. Farrell. Web of influence. *Foreign Policy*, (145):32–41.

[14] J. Fleiss. *Statistical methods for rates and proportions Rates and proportions*. Wiley, 1973.

[15] M. Frank, B. Dong, A. P. Felt, and D. Song. Mining permission request patterns from android and facebook applications. In *ICDM*, pages 870–875. IEEE, 2012.

[16] S. Garfinkel and M. F. Theofanos. A collection of non-breach privacy events, February, 2016.

[17] S. Gibbs. Facebooks privacy policy breaches european law, report finds. *The Guardian*, February 23, 2015.

[18] S. Gibbs. Samsung's voice-recording smart tvs breach privacy law, campaigners claim. *The Guardian*, February 27, 2015.

[19] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7(21):219–222, 2007.

[20] G. Golan. Inter-media agenda setting and global news coverage. *Journalism Studies*, 2006.

[21] S. Greenhouse. Public to get online access to us workplaces' injury and illness records. *The Guardian*, May 11, 2016.

[22] T. Guardian. Guardian open platform. http://open-platform.theguardian.com/. Accessed: 2016-3-3.

[23] A. C. Gunther. The persuasive press inference effects of mass media on perceived public opinion. *Communication Research*, 1998.

[24] E. Harris. Information gain versus gain ratio: A study of split method biases. In *ISAIM*, 2002.

[25] M. Helft. Critics say google invades privacy with new service. *The New York times*, February 12,2010.

[26] C. M. Hoadley, H. Xu, J. J. Lee, and M. B. Rosson. Privacy as information access and illusory control: The case of the facebook news feed privacy outcry. *Electronic Commerce Research and Applications*, 2010.

[27] M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD 2004*, Aug 22-25, 2004.

[28] D. Irvine. Ny times public editor admits paper has a liberal bias. http://www.aim.org/don-irvine-blog/ny-times-public-editor-admits-paper-has-a-liberal-bias-video/, August 21, 2013.

[29] W. Jiang, M. Murugesan, C. Clifton, and L. Si. t-plausibility: Semantic preserving text sanitization. In *ICCSE 2009*. IEEE.

[30] A. C. Johnston and M. Warkentin. Fear appeals and information security behaviors: An empirical study. *MIS Quarterly, Vol. 34, No. 3*, pages 549–566, September 2010. Accessed: 26-02-2016.

[31] S. Kiousis. Explicating media salience: A factor analysis of new york times issue coverage during the 2000 u.s. presidential election. *Journal of Communication*, pages 71–87, 2004.

[32] M. Koppel and I. Shtrimberg. Good news or bad news? let the market decide. In *Computing attitude and affect in text: Theory and applications*, pages 297–301. Springer, 2006.

[33] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer Science & Business Media, 2007.

[34] C. D. e. a. Manning. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[35] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46, 1951.

[36] K. Meissel. A practical guide to using cliffs delta as a measure of effect size where parametric equivalents are inappropriate. In *ACSPRI Social Science Methodology Conference*, 2010.

[37] C. C. Miller. Another try by google to take on facebook. *The New York times*, June 28, 2011.

[38] D. C. Mutz and J. Soss. Reading public opinion: The influence of news coverage on perceptions of public sentiment. *Public Opinion Quarterly*, pages 431–451, 1997.

[39] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in Twitter. In *SemEval*, Atlanta, Jun 2013.

[40] C. Nippert-Eng, M. Carlock, N. Nimchuk, J. Melican, N. Kotamraju, and J. Witte. Privacy and technology: Newspaper coverage from 1985 to 2003. *American Sociological Association Annual Meeting*, 2005.

[41] A. Ortigosa, J. M. Martín, and R. M. Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541, 2014.

[42] P. K. Murukannaiah, J. Staddon, H. Lipford and B. Knijnenburg. PrIncipedia: A privacy incidents encyclopedia. Working paper at the 2016 Privacy Law Scholars Conference.

[43] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[44] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[45] S. T. e. a. Peddinti. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, 2014.

[46] J. Ramos. Using tf-idf to determine word relevance in document queries, 1999.

[47] S. Romanosky. Examining the costs and causes of cyber incidents. In *Twelfth Annual Forum on Financial Information Systems and Cybersecurity: A Public Policy Perspective*, January, 2016.

[48] D. Saez-Trumper, C. Castillo, and M. Lalmas. Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1679–1684. ACM, 2013.

[49] C. Savage. F.b.i. violated rules in obtaining phone records, report says. *The New York Times*, January 20,2010.

[50] A. e. a. Shahbahrami. Simd vectorization of histogram functions. IEEE Computer Society, 2007.

[51] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference*.

[52] K. C. e. a. Smith. Relation between newspaper coverage of tobacco issues and smoking attitudes and behaviour among american teens. *Tobacco Control*, 17(1):17–24, 2008.

[53] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*. Citeseer, 2013.

[54] D. J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review, Vol. 154, No. 3*, January 2006.

[55] S. Soroka, L. Young, and M. Balmas. Bad news or mad news? sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science*, 2015.

[56] J. Staddon, P. Golle, and B. Zimny. Web-based inference detection. In *USENIX Security 2007*.

[57] J. Staddon and A. Swerdlow. Public vs. publicized: Content use trends and privacy expectations. In *HotSec*, 2011.

[58] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006.

[59] M. Trussler and S. Soroka. Consumer demand for cynical and negative news frames. *The International Journal of Press/Politics*, 2014.

[60] T. Vega. A call for a federal office to guide online privacy. *The New York times*, December 16,2010.

[61] A. J. Viera and J. M. Garrett. Understanding interobserver agreement: The kappa statistic. *Fam Med 37, 5 (2005)*, pages 360–363, 2005.

[62] Y. Wang, G. Norice, and L. F. Cranor. Who is concerned about what? a study of american, chinese and indian users privacy concerns on social network sites. In *International Conference on Trust and Trustworthy Computing*, pages 146–153. Springer, 2011.