No (Privacy) News is Good News

An Analysis of Privacy News in the U. S. and U. K. from 2010-2016

Karthik Sheshadri, Nirav Ajmeri, and Jessica Staddon

North Carolina State University

Abstract

News is a popular and influential source of privacy information, and so, an important information source to analyze towards understanding privacy related policy, product development and user perceptions. We provide the first large-scale text mining of privacy news using nearly 1700 articles from the New York Times and the Guardian over the years 2010-2016. Results of four independently-trained sentiment classifiers, show that New York Times privacy news is predominantly negative in sentiment, and more negative than randomly sampled news, and some issues of global concern like the Syrian refugee crisis and the 2015 Paris terrorist attacks. We also provide a frequency analysis of entities involved (e.g. companies) over our entire data set and within certain privacy categories. Our analysis demonstrates that news text can be mined to facilitate both the tracking of important privacy events and efforts to build expressive taxonomies of privacy.

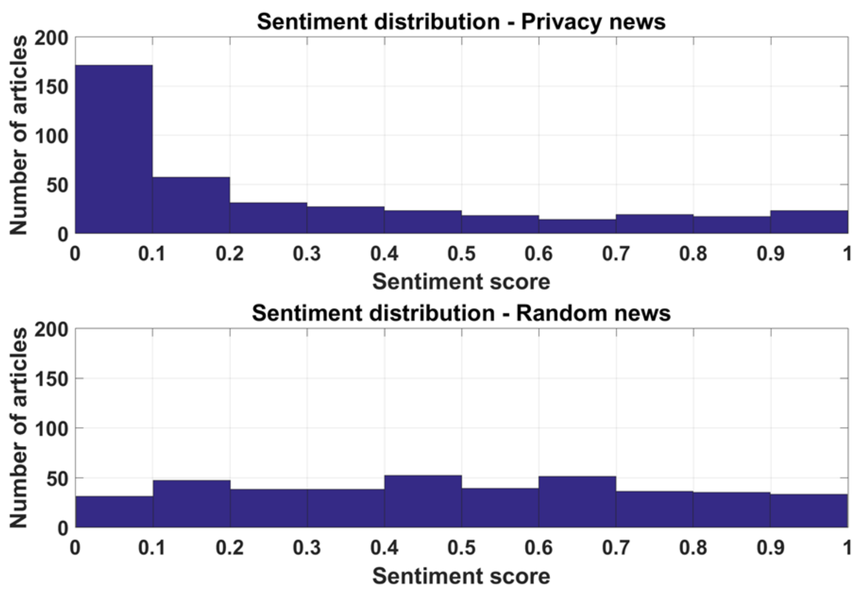


Figure 1: Sentiment score distribution from a Naive Bayes classifier for 400 privacy news articles and 400 randomly selected news articles, all from the New York Times. Measured sentiment increases in positivity as scores increase. Random news articles are almost uniformly distributed across the range of sentiment scores, but privacy news articles are heavily biased towards the negative end of the spectrum.

1 Introduction

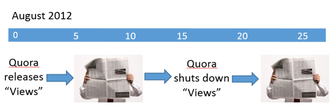


Table 1: Negative sentiment news about Quora's *Views* feature caused *Views* to be shut down. This motivates news as an influential data source within the privacy ecosystem.

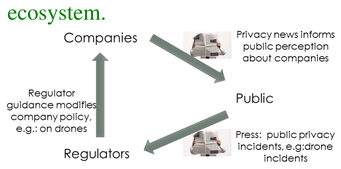


Table 2: News as a communication channel for privacy information. Public perceptions about privacy in general and companies in particular are driven by privacy news publishing.

While news publications are a popular vehicle for communicating privacy information, they are not bystanders to the privacy ecosystem. Rather, news publications are intertwined in the ecosystem in that, for example, privacy incidents are communicated through news, which in turn drives regulation and policy, both of which are themselves often the subject of news articles. Evidence of the intimate nature of privacy news, technology and policy development comes both from regulatory documents and product changes.

We find that privacy news is predominantly negative (Figure 1) and significantly more negative than randomly selected news and events involving significant human suffering like Apartheid2 (Table 3). 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

Data Set	Negative %	Neutral %	Positive %	Number of Articles	p value	Cliff's delta
Privacy	68.91	17.89	13.20	682		
Syrian Refugee Crisis[#]	67.45	21.48	11.04	2854	< 0.05	negligible
China's economy[#]	65.22	22.21	11.59	716	< 0.05	negligible
Paris Attacks[#]	64.16	26.34	9.42	1359	< 0.05	negligible
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
China's one-child policy	45.85	41.89	12.26	1998	< 0.01	medium
Apartheid	37.84	55.40	6.76	2870	< 0.01	medium
Random1	33.4	43	23.6	500	< 0.01	large
Random2	38	41	21	500	< 0.01	large

[#] Stanford's Core NLP parser failed to compute sentiments for a small number of articles in these data sets

Main Findings

1. Privacy news is consistently and significantly more negative than random news, and so is likely to affect public perceptions of privacy.
2. Privacy news is featured more prominently than random news.
3. Framing is more negative than random news at a weakly significant level.
4. Our keyword analysis returns most themes addressed in the surveys of 2010-2015, for example, mobile phone providers, targeted advertising, and so on.
5. Keyword analysis also isolates most privacy incidents in manually curated privacy incident lists.
6. Keyword analysis also isolates most privacy incidents in manually curated privacy incident lists

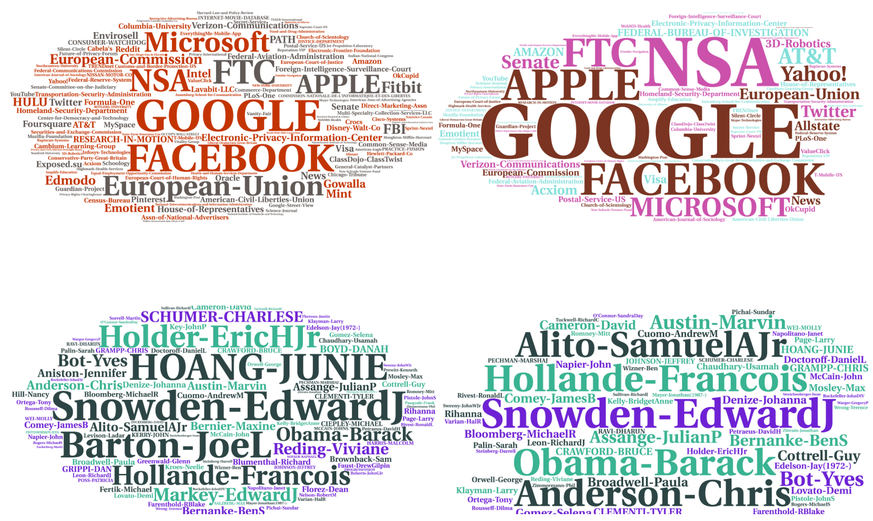


Figure 2: Support for organizations and people in the complete collection of NYT privacy articles (left), and the NYT privacy articles classified as having negative sentiment using the Stanford classifier (right).

Materials and Methods

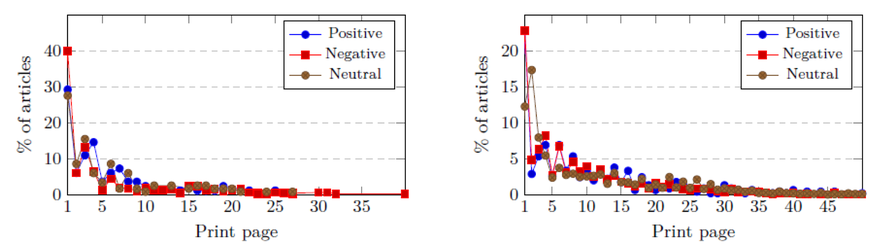
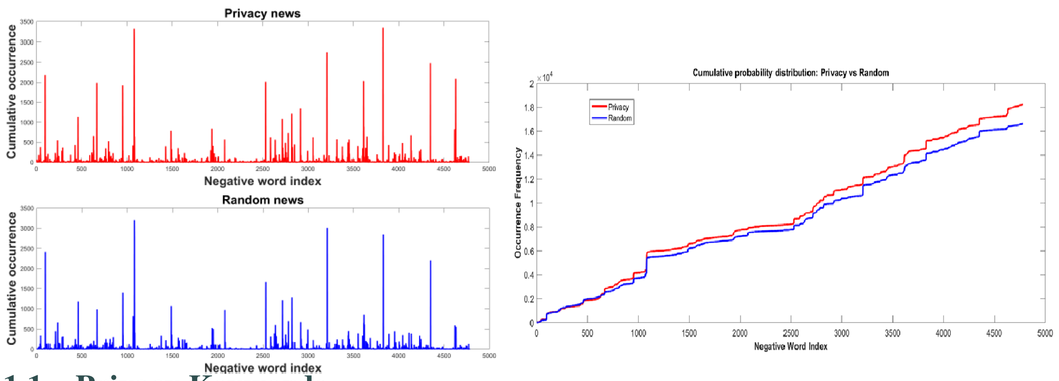


Figure 3: (Left) Sentiment and print page of NYT privacy related articles as measured by the Stanford Core NLP parser. 38 articles of the 682 privacy related articles in NYT are web-only articles or do not have associated page numbers. (Right) Sentiment and print page of NYT non-privacy news articles including coverage of a 2016 US snow storm, Apartheid, Chinas economy, Chinas one child policy, Paris attacks, Syria refugee crisis, and the Zika virus. Sentiment is measured by the Stanford Core NLP parser. A total of 4995 of the 10438 articles are web-only or do not have an associated page number.



1.1 Privacy Keywords

Figure 4: (Left) Relative frequency of occurrence of the 4783 standard negative words in the Guardian privacy news and [illegible]F's of privacy and random framing.

Unigram Keywords	privacy; Facebook; rights; Snowden; 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

Table 3: Discriminative Privacy keywords

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 re ect 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. Up to 218 million Zynga users were affected by the bug.

Use Case: Privacy Incidents Database

Wikipedia for privacy[1]. Privacy News/Incident classifier returns F-measure of 93%. Keyword analysis together with news/incident data sources inform trend analysis for privacy.

[1] https://research.csc.ncsu.edu/privacyincidents/pages/index.php