CS224N: Final Project
Investigating the Web's Dark Matter

Danaë Metaxa-Kakavouli
CS Ph.D. Student
Stanford University
danae@cs.stanford.edu

1 Abstract

People populate the web with content relevant to their lives, content that millions of others rely on for information and guidance. However, the web is not a perfect representation of lived experience: some topics appear in greater proportion online than their true incidence in our population, while others are deflated. This paper presents a large scale data collection study of this phenomenon. We collect webpages about 21 topics of interest capturing roughly 200,000 webpages, and then compare each topic's popularity to representative national surveys as ground truth. We find that rare experiences are inflated on the web (by a median of 7x), while common experiences are deflated (by a median of 0.7x). We further examine topics reflecting opinions and personal views by training supervised machine learning classifiers using crowdsourced page annotations, and find that simple logistic regression classification performs with nearly perfect accuracy on the simpler classification task, but much lower accuracy on more complex classification.

2 Introduction

The web may be the most expansive record of human experience to date, yet there are striking examples of the ways in which we have authored a distorted reflection of our lives [3, 5]. Consider chest pain, for example: if someone browses the web and looks at pages about chest pain, they might conclude that chest pain will signal an imminent heart attack rather than a temporary annoyance [24]. If the web differs significantly from lived experience, then, we must modify our tactics for using the web to inform our research as well as personal decisions.

In this paper, we seek to measure the difference between what people actually experience and what experiences are given visibility on the web. For 21 topics (e.g., religion, smartphone ownership, opinion on same-sex marriage) made up of 74 total components (traits, e.g., being Christian, Jewish, or Muslim; owning an iPhone or Android; supporting same-sex marriage or not) we perform a geographically-restricted web crawl from a neutral seed query. We then annotate a sample of the resulting pages to quantify how the internet represents each component, and compare the proportion to representative ground truth statistics from the same region (e.g., Pew surveys in the United States). We subsequently focus on topics involving political opinion issues, and use the annotated sample of webpages along with those pages' HTML to train two classifiers: the first to decide whether a page is on- or off-topic (i.e. whether it broadly discusses a specific topic), and the second more specific classifier to determine the page's stance (i.e. whether its stance is pro-, anti-, or neutral on the given topic).

We find that uncommon experiences ($< 10\%$ of ground truth) are inflated by a factor of 7 relative to their ground truth value, dominant experiences ($> 60\%$) are deflated to about 0.7 times their ground truth value, and components in the middle region appear roughly in proportion to their ground truth counterparts, at a factor of 0.9. We also find that with simple logistic regression classification techniques we achieve 96.50% accuracy for the on/off-topic classifier, and 75.57% accuracy for the pro/anti/both classifier, both of which are comparable with human accuracy.

In summary, we propose (1) a novel method (crawl-vs-ground-truth) for quantifying the difference between human experiences and what people share on the web. Using this method, we (2) observe that unpopular experiences are overrepresented by nearly 7x on the web and that popular experiences are underrepresented by 0.7x. We then (3) train high accuracy classifiers to allow this analysis to be done automatically at large scale, even across the entire web.

3 Related Work

In many domains, information on the web performs well as a reflection of lived experience. Previous research has shown that online media, for instance, accurately portrays our ground truth friendships [4]. Further, social media data can be used to train

models that effectively model trends in health [20], the stock market [7], movie box-office success [1], and even some elements of our personality profile [10].

In some domains, however, information collected by the web compares less favorably with ground truth reality. Cultural and social biases affecting content creation are well-documented on Wikipedia [3, 5], as well as on social media sites [12]. Search engines, a popular tool for finding information online, are a similarly imperfect representation of the world, as are the queries users issue to them, which can carry subtle personal, social, and political biases [23, 11].

Why does the Web accurately portray some experiences and not others? Looking more broadly at the entire web, there are a variety of reasons content created online might deviate from lived experience. These include individual-level biases and behaviors such as availability bias, diffusion of responsibility, and self-censorship [13]. Such behaviors stem from individual motivations both intrinsic and extrinsic in nature [15, 8], including personal reputation, monetary compensation, altruism, and self-expression [19, 18].

Efforts attempting to correct information imbalances on the web have employed a variety of tactics, including encouraging members of online communities to contribute by emphasizing the uniqueness of their perspectives [2], guiding people to work that needs to be done [6], and by facilitating the consumption of counterattitudinal information through the use of browser widgets [17].

The implications of the web's reflection of human experience are varied and meaningful, particularly because web users do not understand these discrepancies [9], and because individuals can mistake the frequency of encountering an opinion for a proxy of its ground truth frequency [22]. This can have serious consequences for individual web users, as well as for researchers using web data to make predictions about the real world [24, 12]. Understanding the web's ability to reflect or distort reality is critical both for consumers and producers of information online.

4 Crowdsourcing

We begin by looking at how closely the volume of experiences reported on the web match ground truth data. To investigate this relationship we compared content from web crawls to population-representative surveys such as Pew.

To cover a broad sample of topics, we began with three major categories of information:

- *Identity topics*, reflecting affiliations: e.g., religion, political party
- *Experience topics*, describing people's actions: e.g., smartphone ownership, sport viewership
- *Opinion topics*, reflecting personal views: e.g., same-sex marriage, marijuana legalization

Identity topics reflect passive beliefs, experience topics reflect active decisions, and opinion topics reflect personal beliefs.

4.1 Method

We chose 7 topics in each of the above types for investingation, yielding a total of 21 topics (Table 4). Some, such as abortion, are hotly debated; others, such as airline popularity, are less active. Each category had between two and six components ($\mu = 3.5, \sigma = 1.5$), for a total of 74 components.

After identifying the ground truth for each topic (see all ground truth numbers in the Appendix, Table 5), we (i) collect a set of on-topic webpages through a crawl that simulates the averaged behavior of random web users. We then (ii) label a random sub-sample of on-topic webpages according to each topic's components. Finally, we (iii) compare the statistical estimates produced by crowd annotation to nationally-representative survey statistics.

To discuss the crowdworker component more specifically, for each topic, after collecting 10,000 relevant pages, we sent a subset of 500-600 on-topic pages to crowdworkers from Amazon Mechanical Turk, where microtask workers labeled the relevance of each page to the given topic, and to all components within that topic. For example, for the same-sex marriage topic we asked crowdworkers whether the page contained content related to same-sex marriage, and whether the page reflected a pro-same-sex marriage viewpoint, an anti-same-sex marriage viewpoint, or both.

To verify the quality of our annotations, we hand-annotated 20 pages for a subset of 6 topics (2 identity, 2 experience, 2 opinion) blind to Turker annotations, and calculated Cohens kappa as a metric of inter-rater reliability. The average unweighted Cohen's kappa across this subset of 6 topics was 0.784, indicating good agreement between our ratings and the Turkers'. Further, since this metric is unweighted, it does not distinguish between inter-rater judgements that were completely different compared to those in which our annotations differed from Turker annotations by judgment on a single component; therefore, we expect this kappa value to be a lower bound. We thus trusted the Turkers' ratings as accurate.

Annotation by crowdworkers produced 600 annotated pages per topic. We experimented with rating additional pages for several topics, and found that components' proportions were fairly

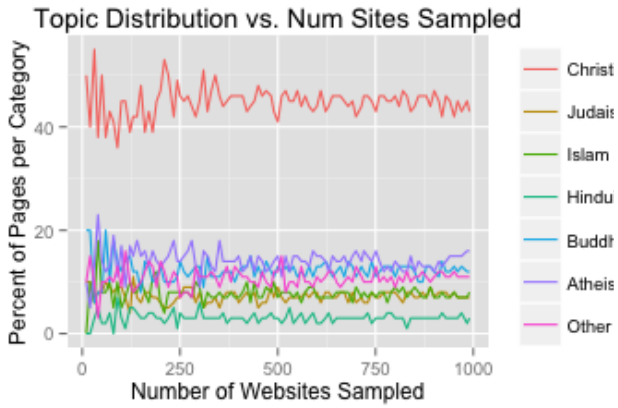

Figure 1: Representation of components within the religion topic stabilized at a sample size of 500-600 pages.

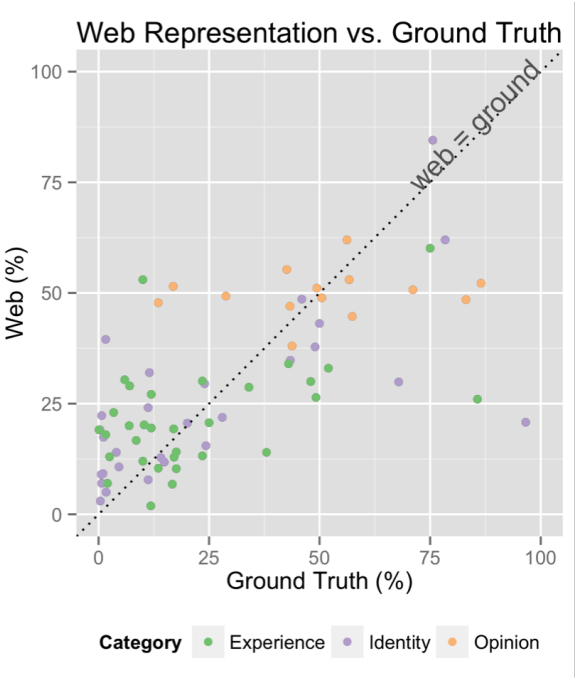

Figure 2: Web crawl vs. ground truth percentages for each of 74 components across 21 topics, colored by topic type.

stable by 600 pages and do not change with more annotated pages (Figure 1).

4.2 Results

In total, our crawler collected approximately 200,000 web pages across the 21 topics, and crowdworkers manually annotated over 12,000 of them. Webpages collected by our crawler ranged broadly in type and content, and included articles from individuals' personal webpages and blogs, news sites (e.g., CNN, New Yorker), organizations' webpages (e.g., Planned Parenthood, Southwest Airlines), and social media content (e.g., specific tweets, Facebook pages).

Figure 2 and Table 5 present each topic components relative ground truth prominence against its representation online. Therefore, in the case of a null result in which every topic component is proportionally represented online and in ground truth, we expect to see all data points fall on the line $y = x$ in Figure 2; the notable deviations from this line in our data reflect discrepancies between the web and reality. Identity and experience topics behaved nearly identically, and opinion topics displayed an even stronger bias (Figure 3).

In our data, *uncommon components*—those with ground truth percentages less than 10%—were over-represented (see Table 4). The median uncommon component appeared online at 6.7 times its ground truth rate. As an example, while only 4% of the American population self-identifies as Atheist or Agnostic, 14% of the webpages crawled on the topic of religion mentioned atheism or agnositicism. Similarly, while heart attacks account for only 10% of chest pain incidents, over half of the pages we crawled (53%) mentioned heart attack as a cause for chest pain.

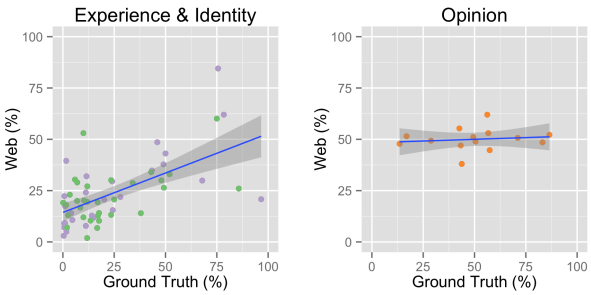

Figure 3: Experience and Identity topics (left) inflate uncommon components and deflate common ones, but components of opinion topics (right) are equally represented.

Dominant components are those whose ground truth is greater than 60%. The median dominant component appeared online at 0.66 times its ground truth rate. A strikingly strong example: while 85% of the American population watches football, only 26% of the rated sports pages discussed it. Likewise, while 68% of the American population is racially white, only 30% of webpages about race mentioned that identity.

Opinion topics deviated from $y = x$ even more strongly, with points clustering near the line $y = 50\%$ (Figure 3). Regardless of ground truth representation, each side of these debates was represented equally relative to the other — for example, opinions about both for and against topics as different in public opinion as same-sex marriage (56%

in favor to 43% against in ground truth) and human cloning (15% in favor to 85% against) both displayed this 50%-50% balance.

As before, we fit a line to opinion datapoints ($web = 0.033 * ground + 48.3$, $R^2 = 0.02$) and tested whether this pattern was significantly different from $y = x$ ($\beta = 1$). We found that opinion topics, like experience and identity, deviated significantly from ground truth ($p < 0.001$).

5 Classifiers

While annotating a random sample of webpages generated by randomly walking a portion of the web does provide significant insight into the information patterns present online, the ultimate goal of this project is to understand the entirety of the web as a single corpus. Having collected crowdsourced annotations for hundreds of pages, we are in a position to automate this process using machine learning classifiers.

For this section, we focus on opinion topics which are unique in their structure—every opinion topic has exactly three components: pro-, anti-, or both. This makes analysis across opinion topics meaningful. Additionally, these were the topics which, in our previous analysis, showed the most dramatic and interesting trend. The opinion topics considered are legality of abortion, morality of human cloning, gun control, morality of polygamy, and mandatory vaccination.

We train two classifiers: an on/off topic classifier, which determines whether a page is broadly relevant to a particular topic (in our crowdsourced tasks this question was phrased "Does the main text of this page discuss [topic]?"); and a pro/anti/both classifier which determines what the main positions of the webpage are (asked as "What specific viewpoints on [topic] are mentioned in the main text of this page?").

5.1 Method

While crawling the web, we collected HTML from the webpages which crowdworkers annotated. As a result, we were able to match 600 pages of text with their crowdworker-annotated labels. We extract text from HTML using the `html2text` library. Preprocessing then continues by tokenizing the text using a Bag of Words model using a count vectorizer into sparse vectors representing the words used on each page. We allow our classifiers to use N-grams of length 1 to 3 (inclusive), and additionally remove common stopwords. We subsequently use the webpage URL to match a page's vector representation to a label.

On/Off Topic Classifier

The on/off topic classifier determines whether a page is broadly "on topic" about a particular topic. The labels for this page are binary, 1 if on-topic, 0 otherwise. Using Python's *Scikit-Learn* library, we train a variety of different classifiers in this stage per topic. In other words, each topic has its own on/off topic classifier. We split our data so that 80% is used in training and 20% is held out for testing. We also perform 10-fold cross-validation on this data.

	LR		**SVM**	
	Test	CV	Test	CV
Abortion	1.0	0.989	1.0	0.979
Cloning	1.0	0.994	1.0	0.993
Guns	1.0	0.992	1.0	0.986
Marijuana	0.996	0.959	1.0	0.932
Polygamy	1.0	0.988	1.0	0.975
Vaccination	1.0	0.985	1.0	0.985

Table 1: Logistic Regression (LR) and Support Vector Machines (SVM) classifiers both perform with high accuracy on both test sets and 10-fold cross-validation (CV) when trained to detect whether a page is on- or off-topic.

	LR		**SVM**	
	Test	CV	Test	CV
Abortion	1.0	0.646	1.0	0.631
Cloning	0.995	0.703	0.995	0.703
Guns	1.0	0.501	1.0	0.520
Marijuana	0.776	0.413	0.768	0.383
Polygamy	1.0	0.606	1.0	0.584
Vaccination	0.989	0.511	0.995	0.568

Table 2: Logistic Regression (LR) and Support Vector Machines (SVM) classifiers overfit significantly and perform much less well when trained to detect whether a page contains pro-, anti-, or both viewpoints on various opinion topics.

Pro/Anti/Both Classifier

This second classifier is a more difficult pass; our inter-rater reliability was 0.784 using Cohen's kappa, indicating good (but certainly not perfect) agreement. This kappa value is an underestimate of pure agreement, since it takes into account agreement occurring by chance.

Here, too, we train separate classifiers for each topic. In this case we use one-vs-rest classifiers since there are three different groups: pro-[topic], anti-[topic], and both topics represented. Again, we use count vectorizers, stripping stopwords and punctuation and allowing n-grams up to $n = 3$. To make the vectors less large, we also ensure that each word has a minimum of two occurrences across the corpus in order to include it.

Figure 4: A word cloud from documents on-topic about abortion reflects relevant words in the debate such as “fetus” and “law”.

Figure 5: A word cloud from documents on-topic about marijuana reflects relevant words in the debate such as “drug” and “legalization”.

5.2 Results

As Table 1 shows, our first classifier is very highly accurate, performing with near perfect accuracy on both test sets and after considering cross-validation. This is reflective of the fact that determining whether a page is broadly on the topic of a certain issue is a fairly simple task; each of these topics have certain distinguishing terms that rarely appear on other pages.

On the other hand, Table 2 reflects the difficulty in accurately identifying what specific viewpoints are present on a page. While these accuracies are much higher than chance (for 3 options chance is $0.3\bar{3}$), they are notably lower than human raters (0.784). This suggests that identifying specific viewpoints and arguments is a difficult task, one which classifiers and humans struggle with alike.

Since these classifiers are based on characteristic words on these webpages, it is also interesting to examine the different common words used in the HTML of our webpages. We can visualize the words used in documents as a word cloud, as shown in Figure 4 and Figure 5.

Figure 6: Differences in word clouds between those who are in favor and those opposed to mandatory vaccination show that the same debate is framed differently by proponents and opponents.

Gun Rights	Gun Control
december	people
2012	2014
people	control
would	suicide
control	violence
think	school
know	news
make	shooting

Table 3: Advocates for gun rights use words including first-person pronouns, whereas those in favor of gun control emphasize words referencing violence.

We can also compare the difference between words in pro- versus anti- documents, rather than grouping all words from a topic’s pages together, both through word clouds and other methods. Figure **??** shows interesting differences in the words used by those who are pro-mandatory vaccination, who reference vaccination successes like “polio,” and those who are against it, and use words like “autism”.

The results, as shown in Table 3, show that opposing sides of a debate use different words when describing the same issue. In the particular issue of gun rights, those advocating for less regulation use first-person verbs like “think,” “make,” and “know.” Meanwhile, those in favor of greater regulation use words referencing school shootings. Notably, dates refering to relevant events or legislation (e.g. December 2012, when the shooting at Sandy Hook occurred, spurring relevant legislation and discussion) are also present among these most common words. For further analysis of text, files with the most common words in each topic and each component as well as their frequencies are included with the handin of this assignment.

6 Discussion

We found that the web overrepresents rare experiences at the expense of the most popular ones, which it underrepresents relative to ground truth. We then show that we can combine crowdsourced labels with webpage HTML to train simple clas-

sifiers that to automatically annotate webpages with relatively high accuracy compared to human raters; however, more complicated classifiation to determine what specific viewpoints are present on a page is not easily learned by a classifier.

6.1 Limitations & Future Work

The need to have human annotations limits our original data collection to 200,000 webpages, with a random sample of only 500-600 pages being annotated. In addition, those pages were collected by randomly walking the web, which mimics web user activity online but requires an unbiased choice of seed for the crawl (otherwise pages collected will skew in some direction and results will not be valid). While in practice we do find that our crawler collects a good balance of webpages, now that we have developed classifiers for automatic annotation we have the opportunity to extend this project across the entire web, for example by running our classifiers across the Internet Archive or CommonCrawl, repositories that take regular snapshots of the whole searchable web.

6.2 Implications

The results from this paper suggest that there are systematic biases in the extent to which different information is represented online. Two types of users are especially affected by these biases: laypersons who rely on the web for information areas they lack expertise, and computational social science researchers that leverage observational data from the Web in their work. Machine learning prediction models, for example, may mistake population priors, and descriptive research from crawls of online communities may misrepresent the actual experiences of individuals in those communities.

Many users that rely on the web are unaware of its shortcomings [9]. As the web is increasingly integrated into everyday lives, individuals' perceptions of the world and their place in it may be skewed by the Web's biases. These skewed contexts can have dramatic results on their behaviors [21].

7 Conclusion

The web is an emergent product of millions of authors. So it is striking that we have, collectively, transformed the relative volumes of our lived experiences so consistently on the web. Through a large-scale web crawl across 21 topics and 74 components, we see that unpopular experiences are overrepresented by nearly 7x on the public web, and popular experiences are underrepresented, at 0.7x. While there are many mechanisms at play, we see evidence that *novelty bias* may capture a micro-scale behavior contributing to the macro-scale result. As a community, we are learning more about *why* people share content on the web (e.g., [16, 18, 19]), but our results make clear that more attention can be paid to naturalistic investigations of *when* and *under what conditions* this occurs (e.g., [14]). With tongue firmly in cheek, we note that the best way to spread this message may be to ensure that it represents a sufficiently uncommon position, and thus achieves a disporportionately large impact.

8 Acknowledgements

Many thanks to Michael Bernstein, who advises this ongoing project, as well as to Gili Rusak and Ella Sung, who were involved previous phases of the project. Thanks also to Chris Manning and the CS224N staff for a great quarter!

References

[1] Asur, S. and Huberman, B.A. Predicting the future with social media. In *Proc. WI-IAT '10*. 2010.

[2] Beenen, G., et al. Using social psychology to motivate contributions to online communities. In *Proc. CSCW '04*. 2004.

[3] Benjamin Hill, A.S. Mapping the wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE*.

[4] boyd, d. Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (editor), *MacArthur Foundation Series on Digital Learning*.

[5] Callahan, E.S. and Herring, S.C. Cultural bias in wikipedia content on famous persons. *JASIS*, 2011.

[6] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. Suggestbot: Using intelligent task routing to help people find work in wikipedia. In *Proc. IUI '07*. 2007.

[7] Gilbert, E. and Karahalios, K. Widespread worry and the stock market. In *Proc. ICWSM '10*. 2010.

[8] Goldhaber, T. Using theories of intrinsic motivation to support ict learning for the ageing population. In *Proc. IUI '12*. 2012.

[9] Graham, L. and Metaxas, P.T. "Of course it's true; I saw it on the internet!": Critical thinking in the internet era. *Commun. ACM*, 46(5):70–75, May 2003.

[10] Graham, L.T. and Gosling, S.D. Can the ambiance of a place be determined by the user profiles of the people who visit it? In *Proc. ICWSM '11*. 2011.

[11] Introna, L.D. and Nissenbaum, H. Shaping the web: Why the politics of search engines matters. *The Information Society*, 16(3):169–185, 2000.

[12] Jungherr, A., Jürgens, P., and Schoen, H. Why the pirate party won the german election of 2009 or the trouble with predictions. *Soc. Sci. Comput. Rev.*, 30(2):229–234, May 2012.

[13] Karau, S.J. and Williams, K. Social loafing: a meta-analytic review and theoretical integration. *J Pers Soc Psychol*, 65(4):681–706, 10 1993.

[14] Kıcıman, E. OMG, i have to tweet that! a study of factors that influence tweet rates. In *Proc. ICWSM '12*. 2012.

[15] Lakhani, K.R. and Wolf, R.G. Why hackers do what they do. *Persp. FLOSS*, 1:3–22, 2005.

[16] Marwick, A. and danah boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society*, September 2010.

[17] Munson, S.A., Lee, S.Y., and Resnick, P. Encouraging reading of diverse political viewpoints with a browser widget. In *ICWSM*. 2013.

[18] Nardi, B.A., Schiano, D.J., Gumbrecht, M., and Swartz, L. Why we blog. *Commun. ACM*, 47(12):41–46, Dec. 2004.

[19] Nov, O. What motivates wikipedians? *Commun. ACM*, 50(11):60–64, Nov. 2007.

[20] Paul, M.J. and Dredze, M. You are what you tweet: Analyzing twitter for public health. In *Proc. ICWSM '11*. 2011.

[21] Prentice, D. and Miller, D. Pluralistic ignorance and alcohol use on campus. *J Pers Soc Psychol*, 64(2):243–56, 2 1993.

[22] Weaver, K., Garcia, S.M., Schwarz, N., and Miller, D.T. Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *J Pers Soc Psychol*, 92(5):821–33, 2007.

[23] White, R. Beliefs and biases in web search. In *Proc. SIGIR '13*. 2013.

[24] White, R.W. and Horvitz, E. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM TOIS*, 27(4):23:1–23:37, Nov. 2009.

Topic Name	Topic Type	Ground Truth	Seed Query	Components
City of residence	Identity	US Census	"US city"	New York, NY; Los Angeles, CA; Chicago, IL; Houston, TX; Philadelphia, PA
Gender	Identity	US Census	"gender"	Female/Women; Male/Men; Transgender/Other
Feminism	Identity	YouGov	"gender equality"	Pro-equality feminist; pro-equality but not feminist
Political affiliation	Identity	Gallup	"American political parties"	Democratic Party; Republican Party; Independent parties
Race	Identity	US Census	"race in America"	White; Black/African American; American Indian/Alaska Native; Asian; Hispanic/Latino
Religion	Identity	Pew	"religion"	Christianity; Atheism/Agnositicism; Judaism; Buddhism; Islam; Hinduism
Sexual orientation	Identity	CDC	"sexual orientation"	Straight/Heterosexual; Gay/Lesbian/Homosexual; Other (Bisexual-Pansexual/Asexual/Queer/etc).
Abortion	Opinion	Pew	"abortion debate"	Anti-abortion; Pro-abortion
Gun ownership	Opinion	Pew	"gun debate"	Anti-gun ownership; Pro-gun ownership
Human cloning	Opinion	Gallup	"cloning debate"	Anti-human cloning; Pro-human cloning
Mandatory vaccination	Opinion	Pew	"vaccine debate"	Anti-mandatory vaccination; Pro-mandatory vaccination
Marijuana legalization	Opinion	Pew	"marijuana legalization debate"	Anti-legalization; Pro-legalization
Polygamy	Opinion	Gallup	"polygamy debate"	Anti-polygamy; Pro-polygamy
Same-sex marriage	Opinion	Gallup	"gay marriage debate"	Anti-same-sex marriage; Pro-same-sex marriage
Airlines	Experience	DOT	"airlines"	American Airlines; Delta; Jet Blue; Southwest; US Airways; United
Chest pain	Experience	[24]	"chest pain"	Heartburn; Heart attack; Indigestion
Fast food	Experience	Rudd Center for Food Policy	"fast food"	Burger King; McDonald's; Subway; Taco Bell; Wendy's
Headache	Experience	[24]	"headache"	Caffeine withdrawal; Tension; Brain Tumor
Music genres	Experience	Statista	"music genre"	Alternative; Country; Hip Hop/Rap; Metal; R&B; Rock
Smartphones	Experience	Pew	"smartphones"	Android; Blackberry; iPhone; Windows
Sports	Experience	Nielsen	"sports"	Baseball; Basketball; Football; Hockey; Soccer

Table 4: Our study covered 74 components in 21 topics, all oriented around human identity, opinions, and experiences. Here we report the source of the ground-truth data for each, as well as the seed query we used for the crawl.

Component	GT %	Web %	Component	GT %	Web %	Component
Religion			**Smartphones**			**Marijuana Legalization**
Christian	78.4	62	Android	48	30	Pro-legalization
Atheist/Agnostic	4.0	14.0	iPhone	43	34	Anti-legalization
Jewish	1.7	5	Blackberry	7	29	**Same-sex Marriage**
Buddhist	0.7	7	Windows	2	7	Pro-same-sex marriage
Muslim	0.6	9.0	**Fast Food**			Anti-same-sex marriage
Hindu	0.4	3.0	McDonald's	49.2	26.4	**Gun Ownership**
Political Parties			Wendy's	16.7	6.8	Pro-gun control
Republican	24	29.5	Subway	11.9	27.1	Anti-gun control
Democrat	28.0	21.9	Burger King	11.9	19.5	**Abortion**
Independent	46.0	48.6	Taco Bell	10.3	20.2	Anti-abortion
Cities			**Headache**			Pro-abortion
New York, NY	43.4	34.8	Caffeine	25	20.7	**Feminist Identity**
Los Angeles, CA	20.1	20.6	Tension	75	60.1	Pro-equality feminist
Chicago, IL	14.1	12.8	Brain Tumor	0.1	19.1	Pro-equality & not feminists
Houston, TX	11.2	24.1	**Chest Pain**			**Sexual Orientation**
Philadelphia, PA	11.2	7.8	Indigestion	38	14	Heterosexual
Race			Heartburn	52	33	Homosexual
White	67.9	29.9	Heart attack	10	53	Bisexual
Black/African American	11.5	32.0	**Music Genre**			Other
American Indian	1.0	9.2	Alternative	17.0	19.3	**Airlines**
Asian	4.6	10.7	Country	13.5	10.4	Delta
Hispanic/Latino	14.9	11.8	Metal	10.0	12.0	JetBlue
Gender			R&B	17.1	12.9	U.S. Airways
Female	50.0	43.1	Rap	8.5	16.7	United
Male	49.0	37.8	Rock	34.0	28.7	Southwest
Transgender	0.3	19.1	**Polygamy**			American
Human Cloning			Pro-polygamy	16.9	51.1	**Sports**
Pro-cloning	13.5	47.8	Anti-polygamy	83.1	48.5	Football
Anti-cloning	86.5	52.2	**Mandatory Vaccination**			Basketball
			Anti-vaccination	28.8	49.3	Baseball
			Pro-vaccination	71.1	50.7	Hockey
						Soccer

Table 5: Ground truth (GT) and web percentages for all 74 components across 21 topics.