Blogs Are Echo Chambers: Blogs Are Echo Chambers

Eric Gilbert, Tony Bergstrom and Karrie Karahalios
University of Illinois at Urbana-Champaign
[egilber2, abergst2, kkarahal]@cs.uiuc.edu

Abstract

In the last decade, blogs have exploded in number, popularity and scope. However, many commentators and researchers speculate that blogs isolate readers in echo chambers, cutting them off from dissenting opinions. Our empirical paper tests this hypothesis. Using a hand-coded sample of over 1,000 comments from 33 of the world's top blogs, we find that agreement outnumbers disagreement in blog comments by more than 3 to 1. However, this ratio depends heavily on a blog's genre, varying between 2 to 1 and 9 to 1. Using these hand-coded blog comments as input, we also show that natural language processing techniques can identify the linguistic markers of agreement. We conclude by applying our empirical and algorithmic findings to practical implications for blogs, and discuss the many questions raised by our work.

1. Introduction

> *It's hardly possible to overstate the value, in the present state of human improvement, of placing human beings in contact with other persons dissimilar to themselves, and with modes of thought and action unlike those with which they are familiar.*
>
> — John Stuart Mill, 1848 [29]

There is no denying the meteoric rise of blogs. Major blogging services launched in 1999; today, the blog index Technorati [43] tracks over 112 million of them. The elite, most heavily trafficked blogs have even started to impact major events in real (offline) life. In 2002, top bloggers led a focused examination of Trent Lott's allegedly racist comments at a political event—shortly afterward, he stepped down as Senate Majority Leader. Many prominent bloggers received official press passes to the 2004 presidential election. In 2004 and 2005, bloggers exposed forged military records shown on *60 Minutes,* leading to the resignation of its anchorman Dan Rather.

While the prominence and power of blogs continue to rise, our empirical knowledge of the blogosphere remains in its early stages. In particular, a number of commentators have questioned the potential of blogs to further fragment the media landscape—effectively shattering it into 112 million pieces. As early as 1996, Nicholas Negroponte theorized about The Daily Me, a newspaper perfectly tailored to your individual tastes and preferences [31]. Nothing appears in The Daily Me to challenge the beliefs you already hold. Cass Sunstein, a law professor at the University of Chicago, hypothesized that blogs may in fact be the modern Daily Me [40, 42]. Building on existing work in group psychology, Sunstein warns that blogs acting as echo chambers could intensely polarize readers and snuff out dissent. Still, the question remains: are blogs echo chambers?

This paper attempts to answer that question. Our empirical study draws on a sample of over 1,000 blog comments made on 33 of the world's top blogs. We focus on comments because they are an essential, and mostly unstudied, aspect of blogs and the "writable" Web. Indeed, on many popular blogs, comments take on a life of their own. One of the blog posts in our sample, a short essay about the press and Barack Obama, triggered 486 comments from readers; those comments occupy more than 80% of the page. In contrast with other work on political linkage patterns in blogs [1, 21], our study covers multiple blog genres and is one of the first to shed light on blog readership [6].

In this paper, we first review relevant results from experimental group psychology and computer-mediated communication. Next we present our methodology for collecting blog comments and the results of our hand-annotation on an *agree–disagree–neither* scale. Using the annotated comments as input, we also show that purely computational approaches can learn the linguistic markers of agreement. We conclude by applying our findings, both empirical and algorithmic, to practical implications for blogs, and outline some of the many questions raised by our study.

2. Literature review

Laying a foundation for how commenters may behave in the blogosphere, we open this section with a short survey of major results from experimental social psychology. We then provide a brief overview of relevant work on computer-mediated communication, blogs and their potential to act as echo chambers. This section concludes with three specific research questions that guide the work presented in the remainder of the paper.

2.1. The psychology of groups: polarization, norms and cascades

Groups dramatically affect individuals' attitudes and behavior—this is perhaps the foundational result of experimental social psychology. For example, groups seeking power engage in riskier behavior than like-minded individuals acting alone, whereas powerful groups act more conservatively [42]. Highly cohesive groups intensely reject deviant individuals [36]. Highly cohesive groups also cause individuals to adopt more extreme versions of their previously held viewpoints [5, 37] and to adopt harsh views of outsiders [38].

Over the last 40 years, many Americans have migrated to highly cohesive communities defined by interests, tastes and political affiliations [7]. Freed from the material obligations that prevented mass migration in the past, Americans seem to have applied the homophily principle of social networks en masse [28]. Commentators and researchers have argued that this homogenizing migration will affect politics and everyday life in dramatically negative ways [7]. Divided into increasingly homogenous communities, Americans have less opportunity to hear the dissent crucial for good societal decisions [20, 41]. Some have proposed social media as an antidote, but others remain skeptical about the medium's potential to overcome groupthink, group polarization and cascades [30, 40, 47].

2.2. CMC, blogs and echo chambers

Researchers have found that the structure of computer-mediated communication (CMC) resembles offline social life, but differs in significant ways [46]. Some claim that CMC levels the social playing field (e.g, race no longer matters); others insist that CMC sometimes reinforces existing offline power relationships [39]. The relative poverty of social cues in CMC may encourage destructive behavior like flaming [24], but others suggest that deeper factors are at work [26].

Blogs swiftly became a major CMC form in the first years of the twenty-first century. The majority of blogs serve as a personal expression forum directed at a small but dedicated audience [22, 23]. However, a small number of blogs have exploded in popularity, reaching audiences that used to be the purview of mainstream media. Blogs differ from traditional media outlets in three crucial ways: rapidly changing content, many offsite links and reader feedback. Commentators have warned that blogs could become, and perhaps already are, concentrated echo chambers in which readers only expose themselves to views they already believe [42]. At the link level, this warning seems prescient. By analyzing the linkage patters of popular political blogs, Adamic [1] and Hargittai [21] independently concluded that political bloggers overwhelmingly link to other bloggers on the same side of the political aisle. At least one prominent liberal blogger, however, argues that political blogs exist expressly for this purpose: to mobilize highly dedicated followers [4].

2.3. Research questions

The literature reviewed above leads us to introduce the following research questions:

R1. Are blogs echo chambers? We examine this question from the perspective of blog comments, and define *echo chamber* as a blog on which more than 64% of the opinionated commenters agree with the blogger. (More on 64% below.)

R2. Does a blog's genre affect its proportions of agreement and disagreement?

R3. Can algorithms learn to detect agreement and disagreement, mainly from blog and comment discourse?

Defining *echo chamber* proves particularly tricky: no existing work has precisely defined it. Since 50–50 is too naïve, we wanted to establish some typical proportions of agreement and disagreement. To do this, we considered face-to-face conversation, and appropriated this baseline in our work on blogs. In an analysis of 75 face-to-face meetings, the authors of [18] report that 18.7% of the time speakers actively agreed or disagreed with one another. Agreements represented 64% of these opinionated moments. So, we call a blog an *echo chamber* when more than 64% of the opinionated commenters agree with the blogger. While the analogy between blogs and meetings is not perfect, we believe it is a reasonable starting place to define *echo chamber.* Blogs offer anonymity not found in face-to-face meetings, but many of the characteristics carry over. Much like blogs, meetings often have a leader, a specific agenda and the participants have a history together.

3. Method

To answer our research questions, we sampled over 1,000 comments from the top 100 blogs as indexed by Technorati [43], a common sampling technique for blog researchers [21, 27]. Technorati's indexing metric, called *authority*, measures the number of distinct blogs linking to each indexed blog over the last six months. Many blog indexes substantially overlap with the

Technorati list [27]. Since the company indexes 112 million blogs, their top 100 list is widely considered to be a list of the most important and influential blogs.

Starting from the top of the list (highest authority), the researchers visited each blog and visually searched for the first post with at least 25 comments. Because of the variability in blog formatting and organization, we could not employ a completely automatic data collection approach. After loading a post in our browser, we used Firebug [15], a Firefox extension, to quickly identify a unique feature for comments on the current blog (e.g., a CSS class name or a DOM tree position). We next used Chickenfoot [9], a Javascript-like programming extension for Firefox, to randomly select at most 30 comments from the post and record them to a file. Our approach ensured that we collected only comments, not other page elements that a completely automatic approach may have identified. We performed our data collection between April 25 and April 27, 2008. Table 1 summarizes the blogs from which we sampled comments, organized by genre. (We did not target particular genres; proceeding down the Technorati list produced these blogs.)

Our data set only captures a thin slice of the blogosphere. This was born out of necessity. As we describe in the next subsection, two human raters assessed more than 1,000 comments and their 33 associated posts. Given limited resources, we felt that it was important to first investigate the most read, most influential blogs. Arguing from existing literature, we hypothesized that a greater proportion of commenters would disagree with an author on blogs that act like public spheres—blogs that give voice to many and are read by many [20]. In this way, influential blogs may in fact provide the most conservative estimate (i.e., an upper bound) of the proportion of disagreement on blogs generally.

3.1. Assessing agreement in blog comments

After reading the 33 associated blog posts, two researchers categorized 1,094 comments into 3 classes: *agree, disagree* and *neither*. While we initially considered a more finely grained scheme (e.g., including *slightly agree, slightly disagree*), our three-category coding scheme allowed us to address our research

Table 1 (right). The blogs from which we sampled comments. *Rank* refers to a blog's position on the Technorati authority index. *Meta blogs* refers to blogs about blogs, blogging and the web. We also report the topic of the post sampled and the total number of comments made on that post. Of these, we chose 30 random comments to construct our data set.

Technology blogs	Post Topic	Comments	Rank
TechCrunch	Twitter	43	2
Gizmodo	Airplanes	56	3
Engadget	DoD	74	4
Kotaku	Tattoos	82	23
Scobelizer	Microsoft	121	30
Gigaom	Facebook	34	35
TUAW	Mac ads	30	37
Joystiq	Wii	53	44
Threat Level	YouTube	62	45
Political blogs			
Huffington Post	Obama	486	1
Daily Kos	Obama	238	11
Think Progress	Scalia	116	26
Crooks & Liars	Voting	132	41
NewsBusters	CNN	67	58
Entertainment blogs			
Boing Boing	Wikipedia	125	5
Gawker	Smoking	144	13
Perez Hilton	Rodriguez	228	20
Valleywag	J. Wales	26	31
Neatorama	Homeless	47	36
Slashfilm	GTA4	58	42
Lifestyle blogs			
Life Hacker	HTML	307	6
Consumerist	Insurance	78	29
uthink	Parenting	167	32
Zenhabits	Love	145	45
Dooce	Lying	336	38
The Sartorialist	The GAP	135	53
Meta blogs			
ReadWriteWeb	Twitter	44	10
Dosh Dosh	Design	157	18
ProBlogger	Workflow	73	21
Copyblogger	Bloggers	61	27
ShoeMoney	Wordpress	54	43
Daily Blog Tips	Twitter	43	34
Matt Cutts	Domains	35	71

First/Second Rater	agree	neither	disagree
agree	**85**	14	1
neither	32	**103**	21
disagree	1	5	**22**

Table 2. Categories assigned by two independent raters working on a random sample of 284 blog comments (roughly 25%) from the entire data set. The raters achieved a Cohen's κ of 0.71 and pointwise agreement of 0.74.

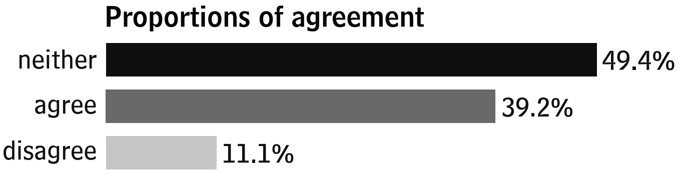

Figure 1. The percentage of 1,094 comments agreeing with the blog author, disagreeing with the author or expressing something else (*neither*). Roughly half of blog comments express either agreement or disagreement, with agreement outnumbering disagreement by 3.5 to 1.

questions without making the experimental design overly complex. The raters focused on the blog author when assessing agreement and disagreement. There is interesting work to be done on assessing inter-commenter agreement. However, we feel the most appropriate place to start this line of work is with the commenter's relationship to the blog author.

Blog comments (and blog posts) exist in highly multidimensional spaces. In this work, we think of our classification scheme as a projection of blog comments onto a one-dimensional plane: *agree* or *disagree*. However, many blog comments only serve to be funny or provide additional information. The *neither* category acts as a destination for blog comments like these, ones that do not fall cleanly into the *agree* and *disagree* categories.

To illustrate what agreement and disagreement actually look like, in response to a post about blogging workflow, an agreeing commenter wrote,

> *Great post and I really like the video. This is extremely similar to the approach I use in writing almost anything ...*

In response to an enthusiastic post about a new Microsoft technology, a disagreeing commenter wrote,

> *Just wait until hackers exploit the print layer to this mesh stuff enough to grab root and start injecting python code ...*

Beginning from an intuitive understanding of agreement, the raters coded 10% of the data and then stopped to check inter-rater reliability via Cohen's κ, a standard measure [10]. In addition to the intuitive understanding of agreement, the raters encountered three delicate circumstances. First, some blog posts only refer (via a link or an embedded image/video) to another blog or news item. Without additional commentary, most such links were simply endorsements of the replicated content. Second, bloggers and commenters often express many opinions and topics in one post. In this case, the raters assessed agreement "on balance." Finally, a handful of the blogs we sampled allowed threaded comments. In this case, the raters carefully read the thread's previous comments to assess whether the commenter referred to the blogger's points or to those of a previous commenter.

After resolving the problematic issues described above (which represented only a small portion of the whole), the raters independently coded the remaining 90% of the data. On a randomly sampled overlap of 284 comments (roughly 25% of the data), the two raters achieved a normalized Cohen's κ of 0.71 and pointwise agreement of 0.74. We report normalized κ, the observed κ divided by the maximum possible κ, because we did not observe balanced categories. In our data set, the *neither* category accounts for roughly half of the cases. Reporting normalized κ is a common practice in this case [12]. Although researchers often debate how to assess a κ value, Landis calls 0.71 "substantial" inter-rater reliability [25] and Altman describes it as "good" [3]. Given the high variance in post topics and styles, plus the inherent fuzziness embodied in the concept of *agreement,* we feel that this is a convincing result.

Table 2 summarizes the categories assigned by the two raters on the overlap set. Most of the discrepancies resulted from the first rater consistently choosing *neither* while the second chose *agree* or *disagree* (row 2 in Table 2). In a post-coding review, it seems that a more finely grained scheme may have helped catch these instances. Only very rarely, in 2 of 284 instances, did the raters assign an agree-disagree pair. This signals a high degree of consistency in the data.

4. Results

Figure 1 illustrates the main result of this section. Roughly half of blog comments take a side on a blogger's post (i.e., *agree* or *disagree*). Of these polarized comments, 77.9% agree with the blog author, 95% confidence interval (0.743, 0.816), using the adapted Wald method [2]. In other terms, 49.4% of commenters take no position, 39.2% agree with the blog author and 11.1% disagree with the blog author. (95% confidence intervals: (0.46, 0.53), (0.36, 0.42), (0.09, 0.13), re-

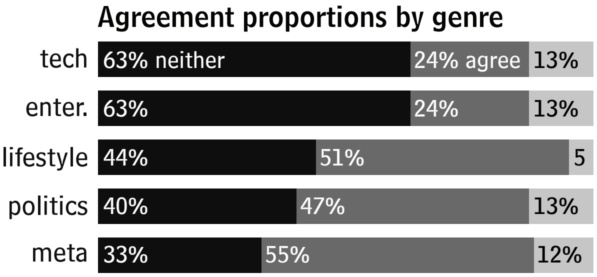

Figure 2. The levels of agreement, disagreement and neither by the five genres in our sample. Blog genre has a significant impact on agreement proportions. Technology and entertainment blogs inspire less polarization and have a much lower agreement to disagreement ratio than the other three genres.

spectively.) 3.5 times as many comments agree with the blog author as disagree.

However, blogs do not uniformly fall into the distribution shown in Figure 1. Blog genre significantly affects the distribution of agreement, $\chi^2(8, N=979)= 86.3$, $p < 0.001$ (see Figure 2). Technology and entertainment blogs cause the least amount of polarization, with only 37% of the comments expressing either agreement or disagreement. Commenters on meta blogs (blogs about blogs), on the other hand, express a definitive position 67% of the time. More broadly, we see two genre groups emerging in our data. The first, comprised of the technology and entertainment blogs, causes little polarization and has an agreement to disagreement ratio of 2:1. The second group, comprised of the lifestyle, politics and meta blogs, polarizes the majority of commenters and has an agreement to disagreement ratio of 9:1.

In order to report our data in the way described above, we had to resolve 74 conflicts between the two raters on the overlap set of blog comments (represented as the off-diagonal entries in Table 2). To do this, we used a random number generator to pick which rater's category would be included in the final set for analysis. Each rater was weighted equally: essentially, this amounted to a coin flip. With regard to the conflicts between raters, the results we report in this section amount to a midpoint, or average, between the two raters' judgements. We have also posted our data on the web for others to analyze, cross-check and perhaps use as input for machine learning algorithms. (We elaborate on machine learning applications of our data in the next section.) The full data set, with all conflicts resolved, and the disaggregated raters' judgements can be found at the following locations:

http://social.cs.uiuc.edu/echo-all.zip
http://social.cs.uiuc.edu/echo-dis.zip

5. Algorithmic prediction of agreement

In this section we introduce our algorithmic approach to predicting agreement. A predictive model could allow readers to quickly assess the actual level of debate across all blogs: think "Technorati for echo chambers." However, this is by no means the only way to attack the echo chamber problem. For example, blog software might be modified to include an agree/ disagree checkbox on comments. This approach may work. A completely automated approach, however, has two main advantages: it requires no additional work from users; and, it potentially applies to all blogs without requiring 112 million software updates.

At a high level, the machine learning and NLP techniques we describe attempt to extract useful information from text to make decisions about it. Text is complex, messy and highly multidimensional. All NLP techniques necessarily aim to build some useful approximation of it. In the following subsections, we describe in detail the features we extracted from our blog data set. Our data set contains a relatively small number of examples by machine learning standards, so we worked hard to explicitly expand the feature space. To do this, we wrote custom text analysis code (in Java and Perl) and used the Weka toolkit [48]. This section concludes by presenting the machine learning algorithms we applied to build a predictive model.

5.1. Lexical features

From each blog comment, we extracted all the unigrams, bigrams and trigrams that occurred at least once in some other blog comment (i.e., at least twice in the corpus). Unigrams are individual words that appear in the text. To illustrate bigrams, consider one the comments we actually encountered,

> *This feels like an echo chamber within an echo chamber!*

From this, we extract the bigrams *this feels, feels like, like an, an echo, echo chamber, chamber within, within an, an echo, chamber !*. Trigrams, similarly, are three-word phrases. All features are lowercased. We included punctuation marks as valid tokens in our lexical features, leading to bigrams like *chamber !*. In various experiments, we found that including punctuation, which is often thrown away, increased classification accuracy by 2–3%. We also experimented with stemming, reducing each word to its root, but it decreased classification accuracy. This may have happened because words like *careful* and *careless* have the same root. Agreement is particularly sensitive to the positive and negative orientations of words. Instead of binary features indicating the presence or absence of a lexical feature, we employed tf-idf term weighting, a standard information retrieval technique [16].

In addition to the n-gram features, we calculated normalized comment length, relative to the other comments from its blog. We also included the percentage of capital letters in a comment (since all features were lowercased) and the position of the comment in a post's comment list.

We made use of the Linguistic Inquiry and Word Count (LIWC) dictionary [33] to extract linguistic aspects of the text we thought would prove relevant. LIWC matches text against pre-compiled lists of word stems assembled into various categories. The lists have been iteratively developed by linguists and psychologists for over a decade and show high validity. For assessing agreement, we looked for the following linguistic categories: numerals, personal pronouns, the word *they* and its variants, past tense verbs, present tense verbs, future tense verbs, question marks, exclamations, negations and swears. (LIWC reports all data as the ratio of matches to the length of the text.)

5.2. Part of speech features

In addition to the lexical features described above, we used a Part-Of-Speech (POS) Tagger to choose the most likely POS tag for each word in every unigram, bigram and trigram. The authors of the tagger report 96% accuracy on a standard test data set [45].

For a lexical feature such as *feels like,* the tagger picks the POS tag VBZ for *feels* (verb, present tense, 3rd person singular) and the tag IN for *like* (preposition or conjunction, subordinating). For every lexical feature, this allowed us to choose not only words, but POS tags, or some combination thereof. Including POS tags led to a 14-fold increase in the number of potential features: 2 options for each unigram, 4 options for each bigram and 8 options for each trigram. For instance, the bigram *feels like* generates the potential features *feels like, feels IN, VBZ like* and *VBZ IN*. After applying the POS tagger to the lexical features described in the previous subsection, we generated 66,231 numeric features describing the blog comment text.

5.3. Named entity features

We also made use of a Named Entity Recognizer to identify references to people and organizations in our blog comments. The authors of the tool report 80–90% accuracy on standard test data sets [14]. In our work, we included two binary features for each comment: one signaling a reference to a person (e.g., *John F. Kennedy*) and one signaling a reference to an organization (e.g., *IBM*).

5.4. Semantic features

We hypothesized that knowing a comment's "on-topic-ness" would help us categorize as *agree, disagree* or *neither*. To measure topic overlap (also called *semantic relatedness*), we computed four metrics. First, for each blog comment and each post we computed its tf-idf vector in the way described above. Two features were derived from it: the tf-idf dot product between a comment and its post and the maximum tf-idf dot product between a comment and all of the post's other comments. A higher score means more topical overlap between the two texts. The score is not normalized to the length of the text, as we wanted to include a feature that did not penalize long posts [19].

Second, we used the tf-idf vectors to compute cosine similarity, a common information retrieval score [16]. We computed it both between a comment and the author's post and relative to all other comments. Cosine similarity can be viewed as a normalized version of the tf-idf dot product feature described above. It measures the angle between two multidimensional tf-idf vectors.

Third, we leveraged WordNet [13] for another angle on the lexical similarity between a blog comment and its associated post [35]. WordNet is a hierarchically arranged lexical database. It allowed us to determine the distance between any two words (e.g., *car* is closer to *bus* than it is to *mouse*). Our feature captures the sum of this distance over all possible word pairs.

Lastly, we borrowed a state-of-the-art semantic technique called Explicit Semantic Analysis (ESA) [17]. ESA uses Wikipedia articles as candidate topics for a snippet of text. Its inventors note that the technique can answer questions like, "How related are 'preparing a manuscript' and 'writing an article'?" Our previously described semantic features fail to capture such subtle relationships. For any given text, ESA generates a list of candidate Wikipedia articles; from it, we calculate the number of articles that overlap between a comment and its associated post.

We knew measuring agreement would contain a significant semantic component. We measured it four different ways simply because we did not know which one would work best. Our first two semantic features captured not only the relationship between a comment and its author's post, but also its relationship to other comments. This perspective allowed us to gauge whether a comment stayed true to the post's topic and also whether it expressed a mainstream viewpoint.

5.5. Sentiment features

We also suspected we would need sentiment analysis to predict agreement. Sentiment analysis predicts the orientation of text along some subjective dimension, e.g., negative–positive, support–reject, love–hate, etc. As coarse measures of sentiment, we collected the following LIWC sentiment categories: affect, positive emotion, negative emotion, anger and assent.

One disadvantage of our data set is its size: 1000 examples is relatively small by machine learning standards. However, a good deal of recent NLP work has focused on sentiment analysis. Using it, we gain access to a larger dictionary and to knowledge about its words and phrases. To this end, we rebuilt the classifiers from [32] and [44]. The authors from [32] and [44] released their data and methods to the research community, allowing us to replicate their work. In [32], the researchers used 1,000 Rotten Tomatoes reviews to build a negative–positive classifier. In [44], the researchers used Congressional floor debates to build a support–reject classifier. The negative–positive classifier achieved nearly 90% accuracy; the support–reject classifier achieved 70% accuracy. We achieved similar results with our rebuilt classifiers. For our features, we used the probabilities along these two spectra.

In addition to the rebuilt classifiers, we leveraged another outside data source: the ABC/Facebook Presidential Debates [11]. In these debates, Facebook asked its users to answer a variety of yes/no questions: “Do you agree with President Bush that the troop surge in Iraq has been working?” When users respond, they annotate their answer with their position: “Yes. If there's anything positive about Iraq, it's the surge.” We manually downloaded 7,000 responses to 7 different debate questions (thereby staying within Facebook’s Terms of Service) and built a similar classifier to the ones described above. Again, our feature is the probability along the agree–disagree scale.

Most NLP systems suffer when moved out of their domains. We did not expect the classifiers described above to perform perfectly, as some of the phrases in them are specific to their domains. (For instance, bloggers rarely use the phrase “water development appropriations.”) We hoped, however, that the classifiers would consistently tilt in a particular direction.

5.6. Blog features

Finally, we included information about the type of blog to which a comment belonged. This amounted to a binary feature for membership in each of the five genres: technology, politics, lifestyle, entertainment and meta. Some blogs belonged to multiple genres. We also included the blog’s Technorati authority score discussed in *Method*.

5.7. Algorithm

To build our predictive model, we used the Bagging meta-algorithm [8] over Complement Naïve Bayes [34]. Other techniques were explored (SVM, boosted decision trees, etc.) but they led to lower accuracy models. Bagging builds multiple versions of an underlying predictor (Complement Naïve Bayes, in this case) and uses plurality vote to make a final prediction. We ran 40 iterations of bagging to build our model. Complement Naïve Bayes addresses many of the shortcomings of traditional multinomial Naïve Bayes, bringing it in line with state-of-the-art algorithms like SVM.

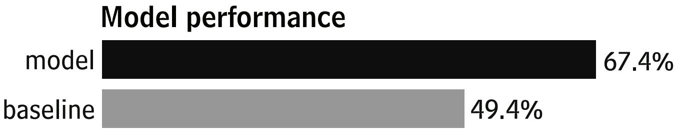

Figure 3. The accuracy of our model for predicting *agree, disagree* and *neither*. *Baseline* refers to predicting the most common category, *neither.* The predictive model we built from Bagging and Complement Naïve Bayes significantly outperforms the baseline. Its accuracy approaches pointwise inter-rater agreement.

6. Algorithmic results

Figure 3 summarizes our model’s performance on the hand-annotated blog comment data set. After constructing the model via the method described above, we evaluated it using 10-fold cross-validation: in ten different trials, we trained the model on a random 90% of the data and tested it on the remaining 10%. The accuracy percentages presented in Figure 3 represent the average accuracy over the 10 trials.

As we mentioned earlier, our classes are not balanced. So, the baseline to beat is the prevalence of the most common class: *neither*, representing 49.4% of the data. (The most naïve model would always pick the most common category.) Also, our model has to learn a three-class problem, a significantly more difficult task than the common binary-class problem. (In other words, even if the model has good information that the comment is not *disagree*, it can still get the prediction wrong.) At 67.4% accuracy, our model significantly outperforms the baseline, $\chi^2(1, N=196) = 7.6, p=0.006$. The model’s accuracy, 67.4%, approaches pointwise inter-rater agreement on the data set, 74%.

Table 3 lists the top 15 features, ranked by information gain, a measure of a feature’s utility. Specifically, information gain computes the effectiveness of splitting a data set into parts based on a particular feature. Its unit is bits. We also report the most likely class given a high value of a feature. In some cases, such as *personal pronouns*, a high value only tells us that the most likely class is not *disagree*. Aggregate measures dominate the top 15 list. This most likely says that given our small data set, the model can estimate better

Top 15 Features	Info Gain	Class Bias
LIWC pos. emotion words	0.079	agree
LIWC affect words	0.049	agree
exclamations	0.043	agree
adjectives	0.041	agree
@	0.041	neither
ellipsis	0.038	disagree
great	0.035	agree
is tech blog	0.034	neither
cosine similarity to post	0.034	not disagree
great [noun]	0.03	agree
personal pronouns	0.028	not disagree
present tense verbs	0.026	neither
[prepos.] [poss. pronoun]	0.026	agree
tf-idf dot product with post	0.026	not neither
coordinating conjunctions	0.026	agree

Table 3. The top 15 features extracted from the blog comments we sampled, ranked by information gain. Aggregate features, not individual words and phrases, dominate the list. *Class Bias* refers to the most likely class given a high value of the feature.

class probabilities for these features (because it has seen more of them). At the same time, our features form a utility distribution with a very long tail. Using just the top 100 features ranked by information gain, our model only achieves an accuracy of 57%. The model achieves the next 10% from features down the tail. This is characteristic of many NLP problems. 15 of the 45 aggregate measures described earlier appear on the top 100 list, including the scores from the movie review classifier, the Congressional classifier and the cosine measure of mainstream viewpoint.

Our predictive model does fairly well on *agree* and *neither*, but performs very poorly on *disagree* (see Table 4). The model achieves a receiver operating characteristic (ROC) area of 0.74 on *neither* and an ROC area 0.77 on *agree*. For *disagree*, the ROC area is 0. The rarity of *disagree* in the data (11%) may be the root cause. We explore this result further in the next section. Also note that Table 4 shows that while the model makes a substantial number of agree–neither/neither–agree mistakes (cells 1-2 and 2-1), it makes the mistakes in a balanced way. We also explore this outcome in the following section.

Confusion Matrix	agree	neither	disagree
agree	**32**	10	0
neither	13	**34**	2
disagree	4	3	**0**

Table 4. The detailed predictions of our model on one of the rounds of cross-validation. The rows of the matrix represent the predicted class, while the columns represent the true class value. The model performs well on *neither* and *agree,* but never correctly detects disagreement.

7. Discussion

While the definition of *echo chamber* in R1 is somewhat tenuous, we feel it is fair to declare that an *agree* to *disagree* ratio of 3.5 to 1 constitutes an echo chamber. Certainly, the result that 77.9% of opinionated commenters agree with the blogger goes meaningfully beyond the 64% mark we appropriated from face-to-face conversation (R1). While we cannot directly compare the face-to-face result with our data (i.e., different experiments and samples), we can confidently say that the overwhelming majority of opinionated commenters agree with the blogger.

Blog genre significantly affects the distribution of agreement (R2). Some blogs (e.g., political and meta blogs) differentially compel commenters to take a side. Further work is needed to establish why. Perhaps there is less to defend or reject on these blogs. Perhaps these blogs draw commenters less interested in supporting or opposing a position.

Our results raise many compelling questions for future research. How much dissent is healthy for a blog? 5%? 50%? If we could reliably measure dissent in blogs, or in social media generally, does it correlate with the blog's success or vitality? Perhaps blog design poorly accommodates conflict. How might we design blogs differently to accommodate it? Does reading highly polarized, highly skewed blogs affect readers in their day-to-day lives, both online and offline?

7.1. Algorithms and echo chambers

We find strong evidence that algorithms can learn the linguistic markers of agreement (R3). In this paper, we demonstrated that natural language processing techniques can predict classes of agreement relatively well, and can significantly beat the prior probability baseline. While we have demonstrated feasibility, the model is only approaching usability. Our model's error rates are most likely too high to accurately assess individual comments. However, given the balanced nature of the errors our model makes, it may be possible to predict a blog's aggregate proportions of *neither, agree* and *disagree*. We discuss how this might look in the next subsection.

Our model's accuracy, 67.4%, is within 2% of the accuracy achieved by state-of-the-art support–reject classifiers [44]. However, extensions may substantially improve our model's accuracy. The first extension, and the simplest, is to collect more training data. Our model learned to predict agreement classes from just 900 examples (90% of the total). If we give our model access to half the data, it classifies with 57% accuracy. The accuracy jump from half the data set to the whole data set is 10%. Collecting 1000–3000 more annotated blog comments, thereby doubling to quadrupling the data set, could bring similar accuracy gains. Adding features that capture social aspects of commenters may also improve accuracy. How many times has the commenter written on this blog? How recently? Does the commenter reveal their identity or remain anonymous?

7.2. Practical implications

Blogs and related social media have attracted many disciples: participatory/collective design, citizen journalism, digital democracy, online politics, collaborative data analysis, etc. We feel that our R1 results bear on these areas. When you put content on a blog, the most likely response will be something like, "that's fantastic…I agree completely…great job!" There is also a deeper cultural issue at work: experimental social psychology suggests that readers in a blog echo chamber will become more polarized and more entrenched in their positions.

We offer our algorithmic work as part of the solution to these problems. Suppose a blogger wants to design for conflict (i.e., obtain a more even agree–disagree distribution) and makes a design change to support it (e.g., prominently linking to an opposing blog). How can she evaluate her effectiveness? We think this is where an algorithmic tool could make a substantial contribution. Using our model, or a slightly enhanced one, the designer could measure conflict before and after the design change.

We also see an opportunity to put our algorithmic work directly in users' hands. We can imagine a metasite that indexes various blogs by their echo chamber measure. Using an echo chamber index, a reader could quickly gauge which blogs generate discussion and which do not. Bloggers that want true discussion and conflict on their blogs could actually incorporate the score into their page design. We feel, for the reasons advanced above, that our model may accommodate a scenario like this right now.

With a small improvement in classification accuracy (most likely the result of more data), we envision a scenario in which a blogger uses a computational model to style and position comments. In the simplest design, opposing columns could hold comments in agreement and comments in disagreement. A blogger might decide to bubble disagreements to the top of the list, or just to be notified when they arrive so as not to miss an opportunity to respond. Many possibilities exist, and to support experimentation we have released our model and its supporting code:

http://social.cs.uiuc.edu/echo.model.zip

7.3. Limitations

This work looks at only the top blogs as indexed by Technorati, over a short time span. While we believe it is a reasonable place to start, we welcome work examining agreement across time and in other blog types, such as diary-style blogs.

8. Conclusion

This paper presents an empirical analysis of blog comments from 33 of the world's top blogs. Agreement overwhelmingly outnumbers disagreement when commenters take a position on a blogger's post. We find that natural language processing techniques can learn the linguistic markers of agreement and, perhaps, be applied toward assessing and redesigning blogs. Our work may raise as many questions as it answers. We look forward to future work examining the theoretical and design questions raised by this paper.

9. References

[1] Adamic, L. and Glance, N. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Proc. LinkKDD,* 2005.

[2] Agresti, A. and Coull, B. A. Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician, 52.*

[3] Altman, D. G. *Practical Statistics for Medical Research.* Chapman & Hall, 1990.

[4] Armstrong, J. and Moulitsas, M. *Crashing the Gate: Netroots, Grassroots, and the Rise of People-Powered Politics.* Chelsea Green, 2006.

[5] Baron, R. S., Hoppe, S. I., et al. Social Corroboration and Opinion Extremity. *Journal of Experimental Social Psychology, 32*(6), 537–560.

[6] Baumer, E., Sueyoshi, M., et al. Exploring the role of the reader in the activity of blogging. *Proc. CHI,* 2008. 1111–1120.

[7] Bishop, B. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart.* Houghton Mifflin, 2008.

[8] Breiman, L. Bagging Predictors. *Machine Learning, 24*(2), 123–140.

[9] Chickenfoot. http://groups.csail.mit.edu/uid/chickenfoot. Accessed June 5, 2008.

[10] Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37–46.

[11] Facebook: US Politics. http://www.facebook.com/politics. Accessed June 5, 2008.

[12] Feinstein, A. R. and Cicchetti, D. V. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

[13] Fellbaum, C. *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

[14] Finkel, J. R., Grenager, T., et al. Incorporating non-local information into information extraction systems by Gibbs sampling. *Proc. ACL,* 2005. 363–370.

[15] Firebug. http://www.getfirebug.com. Accessed June 5, 2008.

[16] Frakes, W. B. and Baeza-Yates, R. *Information Retrieval: Data Structures and Algorithms.* Prentice Hall, 1992.

[17] Gabrilovich, E. and Markovitch, S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proc. IJCAI,* 2007. 6–12.

[18] Galley, M., Mckeown, K., et al. Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies. *Proc. ACL,* 2004.

[19] Glickman, O., Dagan, I., et al. A Probabilistic Classification Approach for Lexical Textual Entailment. *Proc. AAAI,* 2005.

[20] Habermas, J. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgois Society.* The MIT Press, 1991.

[21] Hargittai, E., Gallo, J., et al. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice, 134*(1), 67–86.

[22] Herring, S. C., Scheidt, L. A., et al. Bridging the Gap: A Genre Analysis of Weblogs. *Proc. HICSS,* 2004.

[23] Huffaker, D. A. and Calvert, S. L. Gender, Identity, and Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication, 10*(2).

[24] Kiesler, S., Siegel, J., et al. Social psychological aspects of computer-mediated communication. *American Psychologist, 39*(10), 1123–1134.

[25] Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data.. *Biometrics, 33*(1), 159–174.

[26] Lange, P. G. What is your claim to flame?. *First Monday, 11*(9).

[27] Li, D. and Walejko, G. Splogs And Abandoned Blogs: The perils of sampling bloggers and their blogs. *Information, Communication & Society, 11*(2), 279–296.

[28] Mcpherson, M., Lovin, L. S., et al. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*, 415–444.

[29] Mill, J. S. *Principles of Political Economy.* Boston, 1848.

[30] Mutz, D. C. *Hearing the Other Side: Deliberative versus Participatory Democracy.* Cambridge University Press, 2006.

[31] Negroponte, N. *Being Digital.* Vintage, 1996.

[32] Pang, B., Lee, L., et al. Thumbs up?: sentiment classification using machine learning techniques. *Proc. EMNLP,* 2002. 79–86.

[33] Pennebaker, J. W. and Francis, M. E. *Linguistic Inquiry and Word Count.* Lawrence Erlbaum, 1999.

[34] Rennie, J., Shih, L., et al. Tackling the poor assumptions of Naive Bayes text classifiers. *Proc. ICML,* 2003.

[35] Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proc. IJCAI,* 1995. 448–453.

[36] Schachter, S. Deviation, rejection, and communication.. *Journal of Abnormal Psychology, 46*(2), 190–207.

[37] Sechrist, G. B. and Stangor, C. Perceived consensus influences intergroup behavior and stereotype accessibility.. *Journal of Personality and Social Psychology, 80*(4), 645–654.

[38] Sherif, M. *The Robbers Cave Experiment: Intergroup Conflict and Cooperation.* Wesleyan, 1988.

[39] Spears, R. and Lea, M. Panacea or Panopticon?: The Hidden Power in Computer-Mediated Communication. *Communication Research, 21*(4), 427–459.

[40] Sunstein, C. R. *Republic.com.* Princeton University Press, 2002.

[41] Sunstein, C. R. *Why Societies Need Dissent.* Harvard University Press, 2003.

[42] Sunstein, C. R. *Infotopia: How Many Minds Produce Knowledge.* Oxford University Press, 2006.

[43] Technorati Popular: Top 100 blogs. http://technorati.com/pop/blogs. Accessed June 5, 2008.

[44] Thomas, M., Pang, B., et al. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proc. EMNLP,* 2006., Jul 327–335.

[45] Toutanova, K., Klein, D., et al. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proc. NAACL,* 2003. 173–180.

[46] Wellman, B., Haase, A. Q., et al. The Social Affordances of the Internet for Networked Individualism. *Journal of Computer-Mediated Communication, 8*(3).

[47] Wilhelm, A. *Democracy in the Digital Age: Challenges to Political Life in Cyberspace.* Routledge, 2000.

[48] Witten, I. H. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann, 2005.