A Corpus of Online Discussions for Research into Linguistic Memes

Dayne Freitag  
SRI International  
freitag@ai.sri.com

Ed Chow  
SRI International  
edchow@ai.sri.com

Paul Kalmar  
SRI International  
kalmar@ai.sri.com

Tulay Muezzinoglu  
SRI International  
tulay@ai.sri.com

John Niekrasz  
SRI International  
niekrasz@ai.sri.com

ABSTRACT
We describe a 460-million word corpus of online discussions. The data are collected from public news websites and community-of-interest Internet forums, and are designed to support research on the propagation of socially relevant ideas, a.k.a., “memes.” A structural and statistical description of the corpus is given, and the employed methods of website monitoring, collection, and extraction are described. We also present preliminary linguistic research on the corpus. We show that the corpus represents language from a wide variety of social and psychological communities, that discussion structure and popularity can be predicted in large part from lexical analysis, and that standard epidemiological models provide good fit for diachronic patterns of population-level lexical adoption.

Categories and Subject Descriptors
H.2.4 [Database Management Systems]: Textual databases; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Linguistic processing

General Terms
Corpus analytics, Memetics, Information diffusion

1. INTRODUCTION
Over the relatively short period since its inception, the Web has assumed an increasingly central role in the dissemination of information and the spread of ideas. The widespread adoption of social media, an even more recent phenomenon, has dramatically decreased the friction with which both trivial and momentous ideas spread. In the past, these socially relevant ideas, these memes, might have gained most of their force through official promulgation. In the modern landscape, there is a much stronger bottom-up component to this spread, resulting in more turbulent and, arguably, more interesting patterns in the diffusion of influence.

Information diffusion research has attempted to reconstruct and model these influence patterns, exploiting the computational accessibility of both the linguistic (tweets, blogs, wall posts) and social (friend or follower networks) dimensions of social media. Much of the early work focused on the relatively professional utterances of bloggers, seeking to recover the transmission trajectories of memes (typically URLs or phrases), and to quantify influence as a feature of individual blogs or bloggers [6, 9]. An epidemiological analogy is often applied to the spread of memes through such networks, and models derived or borrowed from this analogy have shown some success in accounting for observed patterns [2, 5]. With the increasing popularity of microblogging, studies of information diffusion in platforms such as Facebook and Twitter have yielded insight into idea propagation and social network formation closer to the “grass roots” [13, 10].

The work described in this paper continues this trend of research away from the professional pundit toward the average citizen, with an emphasis on online discussions. As a rule with a few notable exceptions, Web sites providing news and commentary include comment boards where the reader can respond to specific articles or to other commenters. Such discussions can also be found on special interest Web sites, what we call “communities of interest” (COIs). As will become clear, the harvesting of such discussions poses challenges that have impeded their widespread use in information diffusion and computational linguistics research. But as we also attempt to show, online discussions promise novel socio-linguistic insights.

Our “meme epidemiology” project pursues insights suggested by the epidemiology analogy, attempting to elaborate it in several ways. First, we are interested in what can be learned by treating discussions as “outbreaks.” Like disease outbreaks, discussions grow over time and at different rates. Some become truly huge, while many fail to develop at all. We seek to identify the factors that produce these differences, exploiting the overt connection between an article and the discussion it engenders. Second, we assume that, just as in epidemiology, the energy a discussion or idea exhibits is at least partly attributable to the community or population that hosts it—an assumption that our collection of COI data allows us to test. Finally, we believe that conventional epidemiological models are much more directly applicable to idea diffusion than existing research might suggest. Later, we show that an SIR compartmental model, borrowed with few modifications from epidemiology, accurately models the temporal distribution of lexical expression patterns over several years.

In this paper, we present a 460 million-word corpus of online discussions. We begin in Section 2 by describing the corpus contents and data model. We develop a common vocabulary for corpus ele-
ments, and we provide an overview that reveals some of its salient statistical properties. In Section 3, we present a purpose-built data collection system that has been used to monitor, collect, and extract the data from multiple sources. In particular, our presentation highlights some of the challenges we have encountered in the design of this system and the maintenance of a coherent corpus. Finally, Section 4 presents initial results from three areas of linguistic research being conducted using the corpus: (1) modeling and prediction of discussion structure, (2) linguistic variation between and within website communities, (3) and meme propagation.

2. CORPUS DESCRIPTION

We have been collecting the corpus that is the focus of this paper for nearly a year from public sources. Major features of the collection system and database structure have been stable for approximately six months. We continue to collect data from the sites listed below, and to add to the list of sites.

2.1 Data model

We harvested from 24 distinct websites, each of which have their own way of providing users with the ability to conduct discussions online. As a result, the type and organization of data present on each site may be different from one site to the next. It is therefore necessary as a first step in developing and studying the corpus to develop a common representation for all the linguistic data present—one which generalizes well across the multiple sites and allows for discussion and analysis across the entire corpus.

The data model is centered on discussions as the main representational unit. Discussions are built from two types of discourse unit—articles, which we use to refer to an initial posting of some content (typically a news article or editorial) and comments, which refer to any subsequent statements made in response. Each comment also has an attachment relation linking it either to another comment (when one commenter replies to another) or directly to the initiating article (we refer to this latter type as root attachment). All comments and articles are assigned a posting date (which may include time-of-day information if it is available). Each comment’s author is also obtained, using public user handles when available. The authorship of news articles, in contrast, is not currently available, as we do not have a sufficiently robust mechanism for extracting this information from its embedded position within article text.

We distinguish two main types of websites—community-of-interest (COI) forums and news sites—each type providing certain advantages of interest to the project. COIs explicitly group discussants into more or less culturally homogeneous populations, while news sites make explicit the connection between discussions and the real-world events to which they respond. The two types are distinguished primarily by the way that discussions are initiated (and by whom). For news sites, discussions are initiated by the posting of news articles or editorials that are written by professional authors who are typically not participants in subsequent discussion. Forum discussions, on the other hand, are initiated by discussants themselves, which means the “articles” are usually better described as a discussion “prompt” (though professionally-written articles, or hyperlinks to them, are sometimes posted as articles in forums). News sites and COI forums are also typically distinguished by the nature of their participant community. As the name suggests, COI forums have a more targeted set of common interests, and therefore draw a more focused set of participants.

We are interested in modeling discussions as linguistic objects in their own right, particularly the reply or attachment structure they display, but websites in the corpus often limit certain types of attachment, thus constraining the set of possible discussion threading structures. For example, some sites allow new comments to attach only to the most recently posted comments. In other cases, the recursive depth of the attachment tree is limited. Some web sites eliminate structure altogether, and do not allow comments to attach to other comments at all. These differences limit our ability to generalize some of our findings about discussion structure, but also provide an opportunity to learn about how such constraints affect information propagation. Nonetheless, the applicability of the data model just described is not affected by these differences.

2.2 Descriptive statistics

The corpus consists of approximately 460 million words extracted from 24 websites. A list of the collected websites is shown in Table 1, with those allowing for comment-comment attachment marked with an asterisk (*). As described in the previous section, it is useful to classify the sites into two main types: news sites and community-of-interest (COI) forums. In our selection of COIs, we are interested in choosing sites with a pronounced point of view, while sampling from as broad a range of persuasions as possible.

Table 2 presents summary statistics for each of these two components of the corpus. The data show that comments tend to be longer in COI forums, and that COI forum communities tend to be smaller. Also note that the posting of articles is typical of news sites but not COI forums, though there are some exceptions to this (the COIs richarddawkins.net and vanguardnewsnetwork.com contain posted articles, and some news sites have a few discussions without a posted article).

For many of the websites (typically the news sites), historical data are not made publicly available, so the corpus only contains articles and posts from the period of the collection effort. This means that our archives of such sites contain data spanning periods between 3 and 6 months (depending on when the site was introduced to the collection queue.) Some sites (typically the COI forums) do provide this historical data. For these sites, the collected data spans periods ranging from 1 to 7 years. The website animalssuffering.com has the longest archive and contains data going back to 2004.

An analysis of the distribution of discussion size (i.e., the number of comments in a discussion) reveals interesting properties of the corpus. Namely, we studied the relationship between discussion size and discussion size frequency. In related internet phenomena such as social network connectivity, popularity of websites, or number of email contacts, it has been found that these distributions follow a power-law distribution [1]. But contrary to this pattern, we find that our data require a sub-logarithmic transformation of the two variables (discussion size and discussion size frequency) to produce a linear relationship. This suggests that the stochastic processes that are thought to underly some power-law distributions, such as preferential attachment, may not apply in a straightforward manner to our data [8]. We discuss this further in Section 4.

The data also reveal that temporal factors vary widely across sites. Discussions on the news sites latimes.com and wsj.com, for example, tend to dissipate rapidly, with 95% of comments occurring...
Table 1: A list of collected websites. Those allowing comment–comment attachment are labelled with an asterisk (*).

<table>
<thead>
<tr>
<th>News sites</th>
<th>COI forums</th>
</tr>
</thead>
<tbody>
<tr>
<td>bostonglobe.com</td>
<td>animalrightsdiscussion.com</td>
</tr>
<tr>
<td>foxnews.com*</td>
<td>animalsuffering.com</td>
</tr>
<tr>
<td>huffingtonpost.com*</td>
<td>boston.com</td>
</tr>
<tr>
<td>lasvegasun.com</td>
<td>conservativesforum.com</td>
</tr>
<tr>
<td>latimes.com*</td>
<td>hindudharmaforums.com</td>
</tr>
<tr>
<td>miamiherald.com*</td>
<td>kongregate.com</td>
</tr>
<tr>
<td>motherjones.com*</td>
<td>mothering.com</td>
</tr>
<tr>
<td>npr.org</td>
<td>mpacuk.org</td>
</tr>
<tr>
<td>nymag.com*</td>
<td>richarddawkins.net</td>
</tr>
<tr>
<td>reuters.com</td>
<td>thehighroad.org</td>
</tr>
<tr>
<td>washingtonpost.com*</td>
<td>vanguardnewsnetwork.com</td>
</tr>
<tr>
<td>wsj.com*</td>
<td>vegansoapbox.com</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics for the two main components of the corpus: news websites and community-of-interest discussion forums.

<table>
<thead>
<tr>
<th></th>
<th>News sites</th>
<th>COI forums</th>
</tr>
</thead>
<tbody>
<tr>
<td># of websites</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td># of discussions</td>
<td>148,948</td>
<td>88,551</td>
</tr>
<tr>
<td># of articles</td>
<td>116,449</td>
<td>13,842</td>
</tr>
<tr>
<td># of comments</td>
<td>6,373,186</td>
<td>1,367,586</td>
</tr>
<tr>
<td># of words in articles</td>
<td>53,241,204</td>
<td>6,965,108</td>
</tr>
<tr>
<td># of words in comments</td>
<td>255,267,240</td>
<td>145,414,708</td>
</tr>
<tr>
<td>mean words per comment</td>
<td>40</td>
<td>106</td>
</tr>
<tr>
<td>mean words per article</td>
<td>457</td>
<td>503</td>
</tr>
<tr>
<td>mean unique commenters per site</td>
<td>26,525</td>
<td>5496</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics for the two main components of the corpus: news websites and community-of-interest discussion forums.

3. METHODS OF DATA COLLECTION

Almost all news sites engage their readers by allowing comments to be attached to news articles. In fact, commenting has become so essential that there are now hosted services such as Disqus\(^2\) and Echo\(^3\) offering a comments platform. However, most of these web applications use AJAX technology and require user interaction, making it very difficult to crawl such data [11].

Forum sites that create communities around a specific topic have long been around. Platforms used by such sites are more or less similar. While forum sites can be crawled using classical methods and do not use AJAX, they tend to require registration in order to access forum content.

Our data collection system shown in Figure 1 consists of a discovery module and an extraction module. It is designed to satisfy two major goals: to discover new article URLs from sites of interest, and to extract individual comments and articles. System capabilities include programmatic login and using the Tor\(^4\) network for anonymity.

All twelve of the news sites we harvested, and three COI sites, provide RSS feeds. For these sites, an RSS reader probes feeds for new content, and stores the URLs in a database collection. For sites lacking RSS feeds, or to gather archival data, the process is driven by site-specific configuration files containing seed URLs. A web harvester with an embedded browser starts with these seed URLs, extracts all links and either stores them in a database or queues them and continues navigation. Discovered URLs are then picked up by the extraction module.

The extraction of the actual articles and comments from the HTML pages is the most challenging part: first, AJAX-enabled sites require user interaction with pages to initiate data requests before comments can be navigated; second, each site serves different metadata for comments, preventing the development of a unified data model. We address the first issue with browser-based harvesting, and the second issue with guided extraction and a schema-less document storage.

A generic configuration file used by our system is shown in Figure 2. It consists of sections with key-value pairs, usually expressed in JSON format.

While the meta and login settings are obeyed by both the harvesting and extraction modules, the url-patterns section is mainly used by the harvester. To limit the URL search space, only navigational URLs, such as pagination links, are followed. Links that match article patterns actually point to a main news article or to the head of a thread, and therefore are stored in the database for further processing. The requestRate in the meta section defines the delay between consecutive page requests from a single site; its default value is 15 seconds.

The remaining configuration sections are relevant to the extraction module.

---

\(^2\)http://disqus.com

\(^3\)http://aboutecho.com

\(^4\)https://www.torproject.org/
Our data collection effort faced a number of challenges.

- With some COI sites, a requestRate even greater than the default was needed to avoid detection and subsequent blocking of access.
- Many sites restrict discussion growth along either the time or depth axis, necessitating a fast turnaround between article detection and data extraction. For instance, reuters.com disables commenting after around three days; and foxnews.com refuses to even display comments after three days.
- Our XPath approach to extraction is sensitive to site re-design: during the course of our data collection we experienced one such incident.
- In order to do temporal analysis of the data, dates that the articles and comments are published need normalization. However, every site employs a different format and precision for date stamps: out of twenty sites we identified hundreds of distinct patterns for date formats.
- Especially for sites with threaded commenting, we had to perform extraction over the entire content to determine new comments during our revisits. Previously seen comments were identified by their unique ids within site. These ids are exposed in HTML source to support functionalities such as spam reporting. Content based duplicate detection failed due to changing user signatures.

Our current corpus was collected on hardware with 4x2.66GHz dual-core processors and 32GB RAM, running Linux. All software modules were developed using Python. We chose the document database MongoDB\(^5\) as our storage. We reached to 100GB of database size, including indexes necessary to support crawling. For analysis purposes, we stored snapshots on a local machine and created more elaborate indexes. Firefox was used as an embedded browser and controlled via the Python selenium web driver\(^6\).

### 4. LINGUISTIC ANALYSES

In the previous sections, we described the corpus and the manner of its collection. In the remainder of the paper, we report preliminary research concerning the linguistic and structural patterns associated with meme propogation in online communities.

#### 4.1 Discussion structure

The comments associated with an article form a *discussion tree* with a structure that can vary in shape and size depending on various factors. Learning how and why a discussion grows is helpful in understanding the underlying community and the spread of ideas.

In sites that permit comment–comment attachment, we can observe threadable discussion and present the dependency relationships with other comments on the page. Our data collection system preserves these conversational aspects by identifying the post to which a comment replies, e.g., reply-to (parent and child) relationships.

Our XPath approach to extraction is sensitive to site re-design: during the course of our data collection we experienced one such incident.

In order to do temporal analysis of the data, dates that the articles and comments are published need normalization. However, every site employs a different format and precision for date stamps: out of twenty sites we identified hundreds of distinct patterns for date formats.

Especially for sites with threaded commenting, we had to perform extraction over the entire content to determine new comments during our revisits. Previously seen comments were identified by their unique ids within site. These ids are exposed in HTML source to support functionalities such as spam reporting. Content based duplicate detection failed due to changing user signatures.

---

\(^5\)http://www.mongodb.org

\(^6\)http://seleniumhq.org
Figure 4: Rate of Comment Attachment, by Frequency of Posting, on foxnews.com

Figure 5: Smooth trendlines for Rate of Comment Attachment, by Frequency of Posting

relying to the root article. We have found a split of roughly sixty percent of comments attaching to other comments and forty percent attaching directly to the root node. Moreover, at the author level, we notice a clear relationship with frequency of posting: authors who post more frequently are more likely to attach to other comments as opposed to attaching to the root. Authors who have submitted exactly one comment to a website are much more likely to have replied to the root than to another author. Figure 4 plots the observed rate of attachment to other comments, by number of comments posted, for all authors posting to foxnews.com, along with a smoothed trendline. (The scatter plot is rendered in gray scale, with darker points representing higher conditional probability given number of comments.) Figure 5 demonstrates that this upward trend is characteristic of all sites studied.

What explains this behavior? It is likely that as an individual becomes more familiar with a website community, he or she is more willing to engage other discussants in debate. Individuals who post infrequently are more likely to have recently joined, and may therefore be uncomfortable participating in discussion. By contrast, the 'debater' community will inevitably contain individuals who are unwilling to let others have the last word. Another observation is that among the participants who never bother to reply to—or even read—others' opinions, few are likely to contribute more than one comment to any discussion.

4.2 Predicting Popularity and Attachment

Another of our goals in studying these data is to determine which words, phrases, or other linguistic forms influence the amount of attention the article receives, i.e., its popularity. We describe here a simple experiment with this goal in mind. We formulate the experiment as a machine learning prediction problem involving a regression where the predictors (independent variables) are words in the article text, and the output (dependent variable) is the number of comments the article receives. We study 2,275 discussions from wsj.com occurring in July, August, and September of 2011, training the predictor on 1,820 (80%) of the discussions, and testing it on 455 (20%).

A central challenge to our prediction problem (and many others like it) is that the number of unique word forms in the data far exceeds the number of training samples. (The 2,275 articles in our current dataset contain approximately 60,000 unique word forms.) This means our focus needs to be on limiting the size of our model (regularization) in order to avoid overfitting. Support vector machines (SVMs) are the state-of-the-art in classification and regression in this setting. Interpreting their results, however, proves problematic due to a lack of an easily interpretable relationship between input features and model parameters. We therefore resort to sparse regression techniques and feature selection approaches, a method that has recently proven successful in the context of predicting demographics from social media [4]. Specifically, we focus our effort on an advance in this area called the elastic net [15], which is a regression technique that employs a regularizer that is a linear combination of the $L_2$ norm (ridge regression) and $L_1$ norm (lasso). We use the R package glmnet to fit the model and run our experiments.

We find that elastic-net regression works remarkably well in comparison with SVMs, outperforming it in terms of mean prediction error. After transforming the output variable (number of comments) into a quantile (thus normalizing the output variable to the interval (0,1)), the learned elastic-net regressor achieves a mean error of .176 on the training set. On the other hand, SVM regression achieves a mean error of .264. Analysis of the learned sparse regression shows, for example, that for the selected period of wsj.com, the words obama, taxes, and republicans are the most

7We used the libsvm package with the nu-SVR option. The nu parameter was optimized on the test set by grid search.
effective positive predictors of popularity.

We have also begun to investigate the factors that drive the comment–by–comment growth of discussions, i.e., attachment prediction. Our goal here is to explain why particular branches of a discussion attract different amounts of attention. Viewing the growth of a discussion as a series of attachment decisions, we can ask which of the current nodes (including the original article, or root) is most likely to receive the next reply. We can formalize discussion formation as a generative process, in which each attachment decision is made according to an unknown distribution over existing nodes, and search for distributional models that best account for the observed sequence of attachments. A natural performance metric under these assumptions, borrowed from language modeling, is the attachment perplexity.

Our experimentation in this area involved comparison between three simple models. The first is a baseline uniform attachment model, which considers every comment (as well as the article itself) equally likely to receive the next comment. Second, we considered a preferential attachment model, which assigns each comment a probability that is proportional to the number of comments already attaching to it. This latter model was then refined by incorporating our prior findings about root attachment probability—we assigned a forty percent probability to root attachment, and split the remaining 60% probability among the remaining nodes according to preferential attachment.

The order in which these models are listed above corresponds to a consistent empirical ordering we observe on a range of datasets. Simple preferential attachment yields a considerably lower attachment perplexity than the uniform model, but is further improved by the model that recognizes the special status of root attachment. As we continue work in this area, we are searching for features of the comments themselves—their lexical content, say, or the identity of the commenter—that might allow us to refine further these simple models.

4.3 Community variation and contrast

Because our corpus draws from many different online communities, each with a large collection of authors, there is great potential in our corpus for studying linguistic variation amongst online communities. Additionally, the conversational (and often controversial) nature of the discussions provides a venue for studying contrasting ideologies amongst groups. This section describes some of our preliminary work in this area.

Exploratory analysis of lexical counts in comments shows that the community of commenters within each website in our corpus has remarkably distinct patterns of language use, particularly amongst the COI forum sites. Importantly, the observed distinctions go beyond thematic variation (e.g., differences in topic of discussion), and suggest marked sociological and stylistic contrasts among the sites. For example, we find a large variation in the frequency of personal pronouns, e.g., you, I, and we, which are known to be stable high-frequency indicators of genre [12]. For example, the word I ranges in frequency from 0.95% to 3.75% (a factor of 3.94) across all sites, and the word we ranges in frequency from 0.15% to 0.68% (a factor of 4.59). We also measured vocabulary size by randomly sampling 100,000 words from the comments on each site and counting the number of unique words present in the sample (we report mean results from repeating this procedure 100 times). The resulting figure ranged from approximately 11,500 for the websites motherjones.com, npr.org, and richarddawkins.net, to below 9,000 for mothering.com. Both types of analysis suggest strong distinctions between sites.

By analyzing counts of psychologically-relevant words, the data also suggest distinct psychological characteristics of website communities. In particular, we use a dictionary of word classes distributed with the Linguistic Inquiry and Word Count (LIWC) software program [14]. LIWC is a system that performs psychometric analysis using counts of human-authored (and experimentally validated) word classes such as FAMILIY, POSITIVE EMOTION, and CERTAINTY. Table 3 shows the results of applying a simple LIWC analysis as follows. For a collection of LIWC word classes, we list the website for which the word class has this highest relative frequency. We find these results to match our intuitions about community psychology that have been gained from direct experience with the comments.

### Table 3: Psychometric analysis of websites using LIWC [14]

<table>
<thead>
<tr>
<th>LIWC class</th>
<th>Example</th>
<th>Site with greatest relative freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANXIETY</td>
<td>‘worry’</td>
<td>mothering.com</td>
</tr>
<tr>
<td>FRIENDS</td>
<td>‘buddy’</td>
<td>animalsuffering.com</td>
</tr>
<tr>
<td>SWEARING</td>
<td>‘piss’</td>
<td>vanguardnewsnetwork.com</td>
</tr>
<tr>
<td>CERTAINTY</td>
<td>‘always’</td>
<td>hinduism.org</td>
</tr>
<tr>
<td>DEATH</td>
<td>‘bury’</td>
<td>animalrightsdiscussion.com</td>
</tr>
<tr>
<td>INSIGHT</td>
<td>‘think’</td>
<td>richarddawkins.net</td>
</tr>
<tr>
<td>NEG. EMOTION</td>
<td>‘ugly’</td>
<td>reuters.com</td>
</tr>
<tr>
<td>POS. EMOTION</td>
<td>‘nice’</td>
<td>vegansoapbox.com</td>
</tr>
<tr>
<td>INHIBITION</td>
<td>‘block’</td>
<td>conservativesforum.com</td>
</tr>
</tbody>
</table>

4.3.1 Within-site contrasts

The analyses just described confirm that our corpus covers a wide variety of communities. However, one of the central hypotheses we ultimately wish to test with these data is that coherent but contrasting ideological communities exist within each site. For example, we expect that a controversial site like richarddawkins.net will contain many debates between Darwinians and creationists, and we want to be able to characterize the language use of these two ideological communities.

As a preliminary test of our hypothesis, we perform a simple analysis that contrasts the cooccurrence of three-, two-, and one-word phrases at varying levels of discussion structure. The technique works by measuring how frequently two phrases co-occur in the same discussion and contrasting this with how infrequently they co-occur in the same comment. This allows us to identify pairs of phrases that play opposing roles within conversations. We measure this phenomenon using what we call the bifurcation of two phrases x and y such that

\[
\text{bifurcation}(x, y) = \text{npmi}_{\text{discussions}}(x, y) - \text{npmi}_{\text{comments}}(x, y)
\]

where \(\text{npmi}_{\text{discussions}}(x, y)\) is the normalized pointwise mutual information of the occurrence of phrases x and y in the collection of corpus units

8The word I occurs at least 10,000 times in each of our websites and the word we occurs at least 3,000 times.
specified by \( z \) such that
\[
\text{npmi}_z = \frac{\text{pmi}_z(x,y)}{-\log[\max(p_z(x), p_z(y))]} \\
\text{pmi}_z = p_z(x,y)/p_z(x)p_z(y)
\]
where \( p_z(w) \) is the proportion of units \( z \) in which the phrase \( w \) occurs, and \( p_z(w_1, w_2) \) is the proportion of units \( z \) in which both phrases \( w_1 \) and \( w_2 \) occur.

The results of applying this analysis to our corpus show that some interesting word pairs can be found. From an analysis of latimes.com, for example, we find phrase pairs with high bifurcation such as “tea party movement”—“tea party people.” This pair seems to reflect a positive and negative form of expression for the Tea Party.

It is apparent, however, that our approach warrants some refinement, and the bifurcation analysis also draws out some interesting but unexpected results. For example, the technique reveals phrase pairs like “end of story”—“matter of fact” and “thanks in advance”—“hope that helps,” both of which are indicative of discourse structure rather than opposing ideologies. Also, we find that the technique is very good at distinguishing foreign language comments. Analysis of animalsuffering.com, for example, revealed several discussions in which comments were in both English and French, generating word pairs like “she”—“elle.”

### 4.4 Linguistic meme epidemiology

Our discussion forums provide illustration of memetic outbreaks, in which an idea or attitude propagates throughout a website’s readership. The biological metaphor is apt: a community of susceptible individuals is exposed to an idea expressed by an “infected” individual. Some individuals are “immune” and do not spread the idea, whereas others readily adopt the meme in subsequent posts, becoming propagators of the idea. In discussion forums “exposure” occurs when one individual reads a posting in which an “infected” individual expresses the meme; the contact may or may not result in transmission. As more infected individuals express the meme, the number of contacts and hence infections increases and an “epidemic” ensues. The rate at which contact leads to transmission is dependent on the likelihood that an individual post is viewed by other community members, as well as the attractiveness of the idea being expressed. In many cases the epidemic subsides as infected individuals “recover,” no longer interested in active expression of the meme. The duration of the epidemic is affected by this recovery rate.

We have identified a number of memes that have attained currency during the time periods spanned by our collections. These include pithy epithets such as *Party of No* and catchphrases such as *once great nation*. Our investigations focus on linguistic memes: phrases or lexical entities that can be readily recognized in comments and transmitted with little loss. An example is the family of insult words containing the -tard suffix, such as *libtard* or *religiotard*. Starting in approximately 2007, when this phenomenon was virtually non-existent, the use of -tard as a general-purpose pejorative particle has seen rapid increase in several discussion forums. Our analysis pools all of these forms into a single lexical meme. Another example of a lexical meme is *cretinist*, a derogatory form of the word *creationist*.

Figure 6 depicts the number of authors expressing the -tard lexical meme on the richarddawkins.net site as a function of time. The adoption curve exhibits the classic shape of epidemic growth: rapid initial increase, peak, and gradual decay. Figure 7 illustrates the temporal growth in the number of authors expressing the idiom *kick the can down the road* on the wsj.com site. The difference in time scales for these two epidemics is noteworthy: evidently the *kick the can* phrase went out of fashion relatively quickly.

Statistical and epidemiological techniques can be applied to model meme outbreaks. Using a synchronic approach we might seek factors that predict some measure of severity of an outbreak, or that predict the chance that an individual will be receptive to a particular meme; a diachronic approach might predict the evolution of an outbreak as a function of history.

Our diachronic approach adapts the familiar compartmental models from epidemiology to describe the dynamics of meme adoption. We have found that the classic SIR model yields a qualitatively compelling fit to the observed adoption curve for a number of meme outbreaks: If \( x(t) \), \( z(t) \), and \( w(t) \) denote the number of susceptible, infectious, and recovered individuals at time \( t \), then the growth of these populations is modeled by the following system of equations:

\[
x(t) = -ax(t)z(t) \\
z(t) = bx(t)z(t) - dz(t) \\
w(t) = dz(t)
\]

for parameters \( a, b, \) and \( d \). This classical epidemic behavior is often seen with novelty lexical memes, which tend to enjoy periods of popularity and subsequent decline that are largely unaffected by external events. For example, Figure 8 displays the fit of the SIR model to the tard epidemic observed at six-month intervals beginning in January 2008. Other memes are observed to follow the classic trajectory, or to exhibit a “steady-state” background rate of expression, but later experience a resurgence in popularity because of an external event (e.g., a news item) that heightens the visibility of the meme, and therefore alters the dynamics of adoption.
5. DISCUSSION AND FUTURE WORK

Online discussions provide fertile new ground for linguistic and socio-linguistic research, but the collection of such discussions poses more challenges than comparable social media. As this paper describes, we have worked through these challenges and amassed, over the course of a few months, a corpus of approximately half a billion words.

This data differs from other social media content in ways that open new avenues of investigation. Unlike blogs, there is little or no expectation that online comments will be carefully constructed or even grammatical. Unlike Twitter, discussants are not consciously broadcasting to the world, but are engaging in exchanges with the author of an article or other discussants. Unlike Facebook, discussants face no implicit pressure to maintain an identity. The low barrier to participation, compared to these other forms of social media, makes online discussion arguably more inclusive, contributing to a sample of linguistic utterance from a much broader demographic spectrum. Finally, the author-directed attachment of comments to an article or other comments gives rise to an interesting collaborative multi-document structure, the discussion, which other forms of social media do not provide.

We have only begun to exploit the opportunity this data provides, attempting to account for the spread of ideas, of memes, as an epidemiological phenomenon. None of this paper’s sections offers the final word concerning its respective technical focus. Although our harvesting pipeline has assembled a corpus of considerable size, there are many lingering challenges, such as scaling to a larger number of sites, automating site acquisition and maintenance, and the normalization of comments to account for phenomena such as quoting or excerpting. We have by no means accounted for all the factors responsible for the rate and shape in which discussions grow. We surmise, for example, that different users have different effects on the propensity of a discussion to grow, some of it due to their language, and some to their identity. We see clear linguistic markers for community, but we have yet to measure the influence of community on the spread of ideas. And we have demonstrated the applicability of compartmental models to diachronic lexical adoption, but not more directly to the spread of ideas. All of these objectives remain the focus of future work.

6. ACKNOWLEDGMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL) contract number FA8650-11-C-7119. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

7. REFERENCES


