Going In-Depth: Finding Longform on the Web

Virginia Smith
UC Berkeley
Berkeley, CA
vsmith@berkeley.edu

Miriam Connor
Google Inc.
Mountain View, CA
mkconnor@google.com

Isabelle Stanton
Google Inc.
Mountain View, CA
stanton@google.com

ABSTRACT

tl;dr: Longform articles are extended, in-depth pieces that often serve as feature stories in newspapers and magazines. In this work, we develop a system to automatically identify longform content across the web. Our novel classifier is highly accurate despite huge variation within longform in terms of topic, voice, and editorial taste. It is also scalable and interpretable, requiring a surprisingly small set of features based only on language and parse structures, length, and document interest. We implement our system at scale and use it to identify a corpus of several million longform documents. Using this corpus, we provide the first web-scale study with quantifiable and measurable information on longform, giving new insight into questions posed by the media on the past and current state of this famed literary medium.

Categories and Subject Descriptors

I.5 [Computing Methodologies]: Pattern Recognition;
[Design Methodology]: Classifier Design and Evaluation

1. INTRODUCTION

The *longform article* is a traditional publishing format encompassing a variety of reporting types, from features to investigative journalism. The format is tied together not by content type or writing guidelines, but by a focus on quality and a deep understanding of the subject matter; longform articles are not produced overnight on a deadline. Classic longform pieces include those by authors considered some of the greatest journalists and writers of our time, e.g. “The Duke in his Domain” by Truman Capote [11], “Frank Sinatra Has a Cold” by Gay Talese [52], and “The String Theory” by David Foster Wallace [54].

In this paper, we seek a method to automatically identify longform content. Given longform’s prominence, being able to identify this kind of content enables a variety of applications. These include but are not limited to: building a corpus as the foundation of a web-search system [38]; improving quality or diversity of articles for content recommendation;

providing analytical background content in news applications; or even in creating UI hints for article layout. Here, we demonstrate an additional use of our system, creating a corpus of longform documents for scientific discovery.

In particular, we use our novel system to give clarity to the *the longform debate*: Despite longforms’ prevalence historically, in a time where 140-character tweets are fast becoming the norm, its future is now unclear. Editorial pieces have abounded recently proclaiming that one of three movements is afoot: (1) a death of longform, (2) a renaissance of longform, or (3) a backlash against the genre. Those from group 1 argue that longforms’ relative expense to produce makes it a prime target for cuts, particularly when long-revered publishing institutions are losing classic revenue streams from print advertising [46, 1, 51, 20, 35, 12]. Proponents of the renaissance (group 2) posit that the web is a transformational platform for this style and that its popularity is on the rise [15, 25, 45, 19, 5]. Finally, members of group 3 have mixed opinions, believing that longform *did* gain interest, but that an over-abundance of poorly-produced longform has caused a backlash against the genre [28, 7, 34, 50].

Leveraging our classifier, we provide concrete, quantifiable answers to these questions as they pertain to web publishing. We first develop a robust classifier to identify longform and then, using a massive web crawl with historical signals, aggregate a comprehensive corpus of longform documents on the web. We query this corpus to provide statistical answers regarding the status of longform online, rather than relying on journalists’ opinions. Though prior attempts to quantify the status of longform exist (using hand analysis) [51], a study of this scale has never been performed. This work demonstrates the impact of our classification system, and highlights key techniques needed to carry out similar large-scale studies on the web.

1.1 What is Longform?

Longform is closely aligned with the traditional publishing styles of feature writing and investigative journalism, as well as literary non-fiction and non-fiction storytelling [5]. With the advent of the web, the definition must be broadened to include articles on publications that are web-only, like *qz.com* and *vox.com*, or are self-published on websites like *medium.com* and personal blogs. An initial definition is that longform articles are long, interesting, well-researched, well-written and have a longer ‘shelf-life’ than standard news articles, i.e., they continue to be relevant for months to more than fifty years [5], also known as *evergreen* [26]. We include a few more requirements – that the work must be *original*; *analytical*, i.e., a narrative or commentary, rather than re-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

KDD’15, August 10–13, 2015, Sydney, NSW, Australia.

© 2015 ACM. ISBN 978-1-4503-3664-2/15/08 ...\$15.00.

DOI: <http://dx.doi.org/10.1145/2783258.2788599>.

porting of an event; and aim to *entertain*, i.e., there is a focus on the writing style, and the goal is not just to inform the reader but also to be interesting or thought-provoking.

Longform is perhaps best defined through a series of illustrative examples. For instance, longform includes feature pieces in print magazines on subjects like The New Yorker’s history of elevators [42] or Michael Lewis’s profile of President Obama [30]; investigative journalism like Glenn Greenwald and Ewen MacAskill’s article on the PRISM revelations [22] or a series of 24 reports on Scientology from the Los Angeles Times [48]; as well as feature pieces from online publications like a piece on the history of the Street Fighter movie in Polygon Magazine [43]. One easy way to familiarize yourself with the genre is to peruse a curated longform website like thefeature.net or longform.org.

Longform is not limited to US publishers, with other examples including Canada’s *McLean Magazine* and *The Walrus*; Great Britain’s *The Spectator*, *The Economist* and *Financial Times*; and Australia’s *The Monthly*. Nor is it limited to English, with sources like Germany’s *Der Spiegel* and France’s *Le Monde* and *XXI*. However, it is widely considered that US publishers are the “powerhouse” [5], and we thus focus our study on English language longform.

Longform is also not defined solely by length. For example, a long personal blog post would generally not qualify as it is neither broadly interesting, well-researched, nor evergreen. The vast majority of newspaper articles (regardless of the length) would not qualify as they report the events of the day and are neither analytical nor evergreen. We confirm this empirically (Section 4.3), showing that a classifier from length features alone performs poorly. Finally, we note that the definition of longform cannot be made entirely crisp, as new online sources continue to blur the lines. We ultimately resolve this issue by deferring to human opinion, using human raters to label training data (Section 4.2).

1.2 Contributions

In this paper we make several contributions. First, we give a formal definition of longform and provide a crowdsourcing experiment that can be used to obtain labeled data. Using this data (a training set of less than 16,000 labeled samples), we build a highly accurate classifier of English language longform. This is notable given the huge variety in topic and presentation of longform, and that we develop the classifier using a relatively simple set of features relating only to writing style, word usage, length, and human interaction. The features used by this classifier are publicly available, relying only on access to a web crawl and a natural language syntactic parser. Our use of grammatical language features is surprisingly successful given that previous work using these types of features has focused on more distinct categories, such as distinguishing pieces of news from fiction novels or government documents [49]. Given the breadth and ambiguity of the definition in Section 1.1, it was initially unclear that such a classifier could exist.

While this classifier itself is of independent interest and application, we additionally demonstrate its utility by building a corpus of longform articles to aid in scientific discovery. Using the corpus we provide concrete answers to questions raised by the publishing community. No prior work has been able to systematically create a comprehensive corpus of longform from across the web, and doing so is a testament to the accuracy and scalability of our classification system.

The rest of this paper is organized as follows. In Section 2, we discuss related work on web mining and classification systems, as well as background on longform. In Section 3 we present the novel classifier used to build our corpus, and evaluation of this classifier is discussed in Section 4. Insights and observations gained from the generated corpus are highlighted in Section 5 and we offer conclusions in Section 6.

2. RELATED WORK

Building a substantive corpus of longform has required mining large amounts of web data. The mining of web data to measure cultural, societal, or technological changes is not new. Web mining has been used for a number of diverse applications, from predicting public mood and political uprisings from Twitter data, to detecting cultural differences in geo-tagged images, to measuring linguistic change in online communities [10, 53, 29, 13]. Several researchers have looked specifically at mining web-scale datasets related to news [31, 3, 33, 55, 2]. However, most of these developments were aimed solely at improving news recommendation systems, and none have focused specifically on longform content or on characterizing the nature of this content over time.

We note the importance of the classifier used to create our longform corpus. The use of classifiers to predict specific web content is commonplace throughout applications in document ranking, recommendation, and discovery [8, 21]. To develop our corpus from web-crawled data, we implement a logistic regression classifier, the effectiveness of which has been well-studied and demonstrated by numerous web-mining applications. We use a two-step procedure to develop this model, as described in Section 3.

Lastly, an important subset of features in the classifier leverage natural language processing techniques to detect whether an article’s writing style makes it a longform candidate. Prior work on automatic style and genre detection has used signals derived from syntactic parses and other linguistic cues to identify a document’s style and genre, for example to distinguish between opinion pieces and standard news articles [4, 17, 16] or to identify membership in various online genres (e.g., news, blog posts, and search pages) [47, 49]. Similar techniques have been used to predict whether a piece of writing is well-written and high quality [32, 6].

The Longform Debate.

As discussed in Section 1, there has been much debate recently about the fate of longform, and accordingly, many attempts to investigate the state of the genre. The ubiquity of the web and rise of digital journalism initially caused concern that interest in long content was waning, with notable pieces like Slate’s “You Won’t Finish This Article,” and American Journalism Review’s “The Death of Slow Journalism” [35, 12]. Some argued that, even if interest did exist, the return on investment would not be enough to justify its production cost: This trend was recently documented in studies on the state of news media, indicating that costly investigative reporting styles have undergone major cuts across a number of media outlets [1, 51, 20].

Alternatively, many others have recently declared that longform is undergoing a “renaissance” online [15, 25, 45, 19, 5]. Of these, Arnold’s monograph [5] has the most research-intensive approach, synthesizing accounts from those involved in journalism at all levels, including publishers, writers, editors, teachers, and readers, the ultimate goal being to deter-

mine whether or not longform journalism remains lucrative for media organizations. Her account documents six ‘major-players’ of longform journalism, including Atavist, Seattle’s The Magazine, Longform.org, The Blaze, Byliner, and The Seattle Times. There are additional reports of other “longform startups”, such as Jill Abramson’s offer of up to \$100k advances for longform stories [36]. A history of longform leading to the current purported rise is documented in [37].

Finally, a third camp argues that there is a backlash to the heralded longform renaissance afoot [28, 7, 34, 50]. This side consists mostly of opinion pieces from longform editors and writers, giving nuanced arguments about the hype around longform and importance of quality above quantity. However, some hope to quantify this movement, urging publishers to develop metrics like “dwell time” in order to more accurately measure the popularity of longform [41]. Clearly, it is still undecided as to whether or not longform will maintain its spot as a forerunner in journalistic styles. With a corpus of more than 5.2 million longform documents in hand, this is exactly the question we aim to illuminate.

3. BUILDING A LONGFORM CLASSIFIER

In this section we describe our two-step classifier for identifying longform content. Such a classifier is useful for a variety of applications, including search systems, content recommendation, news applications, or in aiding UI development. Using a classifier is necessary due to the scale of the problem (i.e., identifying a subset of the web), as well as the importance of objectivity and standardization in ratings. We aim to develop a classifier that is simple, transparent, and accurate. These practical design traits ensure that our classification system will be scalable and reliable, as well as easy to maintain. Through extensive testing, model selection, and feature-engineering, we have developed a classifier that exactly meets these needs.

Figure 1: Workflow for developing and using the longform classifier. Classification includes using the NLP classifier as part of the full classifier in a two-step procedure.

Workflow.

Our workflow for developing and using our classifier is shown in Figure 1. We begin by sampling a training set representative of the diverse documents on the web, a task made difficult by the huge number of possibly irrelevant pages (Section 4.1). We then label this set using a crowd-sourced rater system (Section 4.1). Next, we extract a number of features from the documents, which are described in detail in Section 3.1. The features relate broadly to categories of length, document interest, and document quality.

Two-Step Classification.

The extracted features are utilized in a *two-step* procedure. In particular, we construct two classifiers, which we will refer to as the *full classifier* and *NLP classifier* (or natural language processing classifier). The features of the NLP classifier are based on grammatical parse structures from a natural language parser, described in Section 3.1. We use predicted output of the NLP classifier as a feature source in the full classifier. This two-step procedure allows us to separately tune and adjust the features related specifically to natural language, and has the added benefits of reducing the dimensionality of the final model, helping to prevent over-fitting, and mitigating effects of correlation amongst variables. The two classifiers are trained independently on separate datasets, making this distinct from related work on multilevel, nested, or hierarchical models that consider estimation on the full model [56, 44].

Both the full classifier and NLP classifier employ logistic regression, a probabilistic classification model for predicting binary variables $y^{(1)}, \dots, y^{(n)} \in \{0, 1\}^n$ from features $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^d$, with success probability:

$$p(y^{(i)} = 1 | \beta_0, \beta, x^{(i)}) = \frac{1}{1 + \exp(-\beta_0 - \beta^T x^{(i)})}$$

The true values are predicted according to the discriminant function, a linear combination of the input features $\mathbf{x}_j^{(i)}$:

$$\beta_0 + \sum_{j=1}^d \beta_j x_j^{(i)}$$

Finally, in order to generate a corpus of longform documents, we utilize the classifier on a massive set of web crawls, passing the features to the fully trained model in order to determine each document’s final score (Figure 1b). By thresholding this score, a value between 0 and 1, we restrict our web-scale dataset to a corpus containing only longform documents as identified by our classifier. The observations we generate from this final corpus are given in Section 5.

3.1 Longform Features

We look for three major cues in categorizing longform: document length, interest, and quality. These are measured by calculating repeated structure counts, anchor count and quality, and document language and structure, respectively. We collect the following features:

- **Article length:** We use a state-of-the-art system for parsing HTML structure and identifying the main article content from other parts of the page, like comments or the boilerplate. We find the number of characters in the article and bucket this feature into 4 groups: short (<5,000 characters), medium (5,000–10,000), long (10,000–20,000), and very long (>20,000).
- **Anchor:** The number of *off-domain* anchors, as well as the presence of recently occurring anchors (i.e., those newly observed within the last three months). These measure interest in the document from external sources.
- **NLP Classifier score:** A measure of the quality of the document, classifying longform articles based on language and grammatical structure. This feature is bucketed into three buckets: high score, low score, and score aggregated over all articles from a site.

Including the intercept term, this amounts to a total feature size of 10, using 4 features for article length, 2 for anchors, and 3 for NLP classification. However, the NLP classifier itself (described below) uses many thousands of features.

NLP Classifier.

Beyond the length of a document and the interest level of its readers, the measure of a successful longform article lies in the content of the article itself, and particularly in the writing style. A detailed news report on a breaking topic might be long and frequently cited, but will lack the inventive, entertaining, and analytical language definitive of the longform genre. We therefore need a signal for our full classifier that indicates the quality of the article text itself. To capture this, we train a separate classifier on natural language features and use the output of this classifier as a feature to the longform classifier. Like the full classifier, this classifier is a logistic regression model [9].

To train the NLP classifier, we extract natural language signals from the text of each web page. We obtain syntactic parses of the document text with a shift-reduce transition-based parser [39] using standard features and a linear kernel SVM for classification. The parser is trained on dependency parse trees that are generated by applying the Stanford converter [14] (version 1.6.2) to the OntoNotes treebank data [24]. We describe our natural language features below.

- **Token ngrams:** The most common two-, three-, and four-word sequences of tokens in the page text.
- **Parse tree ngrams:** We run our dependency parser over the page text and extract ngrams by traversing the parse tree. For example, the sentence “the furry dog barked” would generate the ngram: *determiner* → *nominal subject* → *sentence root*.
- **Adjacent parse label ngrams:** The most common sequences of parse labels from adjacent tokens, for ngram lengths three and four. An adjacent parse label of length three in “the furry dog barked” would be *determiner* → *adjectival modifier* → *nominal subject*.
- **Part-of-speech ngrams:** The most common sequences of part-of-speech tags from adjacent tokens, for ngram lengths two and three.
- **Sentence length:** Average length of sentences on page.
- **Pronoun person distribution:** The relative frequency of first-, second-, and third-person pronouns.
- **Punctuation frequency:** The relative frequency of different punctuation mark tokens in the text.
- **Delimiting punctuation frequency:** As above, but only considering sentence-final punctuation marks.

As we use multiple ngram-based features, the feature space for the NLP classifier is quite large. We use feature hashing to reduce the feature space and to add regularization [18].

Intuitively, the NLP classifier on its own should not be sufficient to identify longform articles. It does not, for example, capture the veracity of an article or whether its topic is broadly interesting—the anchor features may be better indicators of this. We confirm this intuition experimentally by evaluating the quality of the NLP classifier independent of the full classifier in Section 4.3.

4. EVALUATION

In this section we evaluate the classifier proposed in Section 3. We validate the performance in a number of ways, including thoughtfully constructing training and testing data (Section 4.1); using feedback from crowd-sourced human raters (Section 4.2); and assessing the classifier with a number of statistical metrics (Section 4.3).

4.1 Challenges in Generating Training Data

When dealing with a corpus the size of the web, finding good representative positive and negative training examples is both incredibly important and quite challenging. In this work, we must identify a subset of the web ordering millions of pages from a corpus of trillions. Techniques like random sampling will both find no positive examples, and find relatively “easy” negative examples unrelated to journalism, leading to poor generalization during learning. Instead, we craft our training examples based on intuition for the problem and then show through evaluation that this approach produces a classifier that generalizes well.

Our labeled training data consists of 16,000 examples. For negative examples, we include a random sample of 6,000 articles consisting of at least 1,500 characters from a large collection of news articles. These documents are all given a negative label, with knowledge that the vast majority of news articles do not qualify as longform. We include the length requirement to reduce training bias and ensure that that feature alone is not used in identifying longform (we demonstrate in Section 4.3 that a length-only classifier performs very poorly). In addition, we take a sample of documents that ranked highly in Google’s In-Depth Article Search system, an independent ranking service that uses a host of signals orthogonal to those in our classifier [38]. Using these candidate articles, we have a set of in-house, well-trained human raters label the examples, a process which is described in detail in Section 4.2. From this second set of experiments, we obtain an additional 8,000 positive examples and 2,000 negative examples.

4.2 Human Ratings

Longform is surprisingly difficult to define. While it is easy to identify stellar examples of the genre, and non-examples in the form of news articles, there is a wide range of articles in-between that can only be labeled through human judgement. To evaluate the full quality of our classifiers across this spectrum, we use human raters to evaluate articles based on the criteria defined in the rater prompt (Section 4.2.1). We use this rater task to both generate training data as well as evaluate the quality of classifier scores with additional human judgements.

The answer to each question translates into ratings between 0 and 4. Ratings from each rater can then be averaged for each article in order to determine a single numeric score. In our evaluation, this average score is scaled to between 0 and 1. While we are only interested in the final question about in-depth, we found that constructing the rating task with the first four questions resulted in higher accuracy from the raters as they were forced to consider each component of longform before generating their final rating. An article with an average unnormalized score less than 1 for in-depth is considered to have a clear negative label, while an article with average score of at least 3 is considered to have a clear positive label.

4.2.1 Rater Prompt

Instructions: In this task, you will be asked to rate an “in-depth article”. Please carefully read and consider the description below before completing the task. For the this task, an “in-depth article” has the following properties:

- The content is a text article with the intent to provide information to the user, as opposed to helping the user perform a task or navigate to content.
- The content is interesting to consume while also being informative. Reference materials primarily used to consult or look up facts are not in-depth articles.
- The content has a narrative form. A user would typically read the content from beginning to end, as opposed to skipping to a section of interest.
- The landing page is from a high quality, authoritative source.
- The content is long and detailed enough to be more than just a general summary of a topic. A user who is already familiar with the general subject area would find the content interesting.
- The content has original opinions, research, or synthesis. It is more than a statement of facts that could be found elsewhere.
- The content has broader relevance than a typical news story and would still be interesting to a reader many months from now. For example, a biography or profile of an athlete would meet this criterion but a recap of that athlete’s most recent game would not.

Is the content from an authoritative source?

[Yes - The author or website is known for its high quality journalism and/or expertise in the main content’s subject matter.] [Mostly] [Somewhat - The author or website is not particularly renowned and/or has expertise in a different subject.] [A Little] [No] [Can’t Tell]

Does the content include detailed, original content?

[Yes - Substantial effort went into researching or creating the content. It has information that is not readily available on other websites.] [Mostly] [Somewhat - The content is original but brief, or it is detailed but a summary of facts available elsewhere.] [A Little] [No - Someone else could easily create a webpage that achieves the same purpose as the content.] [Can’t Tell]

Look for the date the content was published and consider how the content relates to events at that point in time. About how long would it take for the article to become substantially less relevant, compared to when it was first posted?

[Never or Almost never - The content will most likely be permanently relevant.] [Several Years] [Three months to a year - The content does not directly relate to current events but may become less relevant over time.] [Up to three months] [Days - The content is “newsy” or about current events.] [Can’t Tell]

Is the content interesting to a user today? Is the content currently relevant and not stale (regardless of its age)?

[Yes - The content is on-topic and currently relevant.] [Mostly] [Somewhat - The content is somewhat stale, but still provides interesting context on the subject.] [A Little] [No - The content is outdated.] [Can’t Tell]

Overall, is the content of the landing page an “in-depth article”?

[Yes - The content has all of the characteristics of an “in-depth article”.] [Mostly] [Somewhat - The content has most of the characteristics of an “in-depth article”.] [A Little] [No - The content has few or none of the characteristics of an “in-depth article”.] [Can’t Tell - Could not view content (Page did not load, or content behind a pay wall).]

4.3 Performance of Classifiers

We evaluate the performance of the classifiers on both holdout training data as well as unseen generalization data, using a number of standard statistical metrics.

Performance on Holdout Data.

Figure 2 shows ROC curves with respect to data held out from our training set, for the full longform classifier (red, solid); NLP classifier (green, dotted); and a classifier trained from only length features (blue, dashed). We train the models using a random sample of 80% of the data, with the holdout set containing the remaining 20%. Numerical results are given in Table 1.

Length features alone do not provide enough performance to warrant using as a classifier, receiving low accuracy and poor AUC results. Though the NLP classifier has a high precision value of .8580, the recall for the NLP classifier is quite low. This discrepancy emphasizes the need to combine the natural language signal with our other signals in order to achieve both high-quality and high-coverage identification of longform articles. We see that the low recall of the NLP classifier is mitigated by the full longform classifier, which achieves precision .9138 and recall .8690.

Figure 2: ROC Curves for the full classifier and NLP classifier on the hold-out set from the training data.

Table 1: Performance of Classifiers on Holdout Data.

	Length Features Only	NLP Classifier	Longform Classifier
AUC	.5782	.8319	.9424
Accuracy	.6128	.6202	.8783
Precision	.7317	.8580	.9138
Recall	.6870	.4031	.8690

Performance on Unseen Generalization Data.

Next, we generalize to a set of randomly selected, previously unseen webpages, evaluating how well the scores generated by the classifiers correlate with human ratings. To guarantee an accurate evaluation, regardless of the distribution of the classifier scores on the articles, we use a stratified sampling method with a reservoir sampler to generate an equal number of articles for each classifier score bucket, i.e. between 0 and 0.1, 0.1 and 0.2, etc. These are then rated using the experiment in Section 4.2. We take the average of the rated scores for each data point, and scale this number to match the output of the classifiers between 0 and 1.

Figure 3: Generalization Results for the Longform and NLP Classifiers.

In Figure 3a, we show the correlation between the average rater score and the score given by the full classifier, for a new sample of 1,000 previously unrated webpages. Figure 3b shows analogous results for the NLP classifier. Even after considering the sampling, it is evident that the scores are somewhat stratified, which is a product of a number of integer-valued features limiting the dimensionality of the space. Comparing these plots, we see that the full classifier is able to more accurately match the scores given by the raters. The optimal correlation ($y=x$) is shown in the black, dotted line, and the results from the full classifier (the solid line in Figure 3a) closely match this. The full classifier achieves an adjusted R^2 value of 0.34, as opposed to the NLP classifier, which has an R^2 value of 0.06.

It is important to remember when viewing these results that the average rater scores are not an ideal predictor as there are discrepancies amongst the raters themselves: If humans can't perfectly agree, it seems unrealistic to expect the classifier to do so. The variation in rater scores is particularly high when answers to the questions were in the middle range, i.e. were rated 1-3 instead of 0 or 4. For documents that were rated on average between 1 – 3, the average deviation from the mean for scores was $\sigma = 1.57$. This is in contrast to scores with average ratings of < 1 or > 3, which had a standard deviation less than .5. Thus, another metric to consider is whether the classifier performs well in areas of high confidence. To investigate this, we additionally consider the performance of the classifier only on data points where the raters are very confident. The classification results of this are shown in an ROC curve, Figure 3c. We see that the full classifier performs quite well in these areas of high confidence, achieving an AUC of 0.91 as compared to 0.75 for the NLP classifier.

Finally, we briefly discuss the relative weights of the features of the trained full classifier, as shown in Table 2. The three specified qualities of longform (length, popularity, and quality) each have a significant impact on the classifier, with feature weights spread roughly uniformly amongst each feature type. In terms of length, articles that are very short are given a negative score, as well as articles that are quite long (e.g., a novella or online book should not be considered an article). The NLP classifier also has a large impact on

determining the outcome of longformed-ness, with low NLP scores negatively impacting the outcome, and high scores for the article or site giving positive impact. The popularity of the page also plays an important role in determining this outcome, with both the number of recent anchors and the number of off-domain anchors being positive indicators.

Table 2: Classifier Feature Weights.

Feature Type	Feature	Weight
	intercept	-0.8795
length	short	-1.9357
	medium	-0.2873
	long	-0.0620
	very long	-0.7678
quality	low NLP score	-1.2197
	high NLP score	0.8067
	site-based NLP score	1.2012
popularity	existence of recent anchors	1.2012
	# of off-domain anchors	1.5936

5. A STUDY OF LONGFORM ON THE WEB

In this section, we demonstrate the utility of our classification system by using it to create a corpus of over 5.2 million longform documents on the web. Using this corpus, we provide concrete numbers that address editorials on the current and future state of longform, i.e., is it dying, or facing a new boom? If there is a boom, are people sick of the trend? For reference, the three dominant arguments are as follows: that there is (1) a decline of longform, (2) a renaissance of longform, or (3) a backlash against the genre.

Surprisingly, despite the seeming opposition of these statements, our analysis of the longform corpus provides evidence for all three camps, as we describe below. Longform is not dying; we see a steady increase in the amount of longform published on the web over the years. However, news publishing on the web is increasing at a faster rate, and so the relative amount of longform published *is* dropping. With

regards to a “longform renaissance,” we see that longform appears to be increasingly popular: Interest in longform relative to news is growing and more sites than ever before are producing longform content. As for the backlash, we do not find evidence to directly support this hypothesis, but we do see a significant spike in the number of new sites producing longform. This provides evidence that longform is a ballooning trend, which forms the basic assumption required for a backlash to develop.

5.1 Building a Longform Corpus

Though we have developed a high quality classifier for distinguishing English language longform articles, additional finesse is required to build a corpus from a web crawl. In order to reduce the space of possible documents, we begin building our corpus by using two additional systems: one for identifying webpages with commercial intent (e.g. Amazon product pages), and one for attributing an author to webpages [27]. We use these systems to filter out clear negative examples to improve the efficiency and quality of our corpus.

The next challenge faced is multi-page articles. In physical print it is necessary that longform articles be paginated, and this tradition is sometimes carried onto the web, making it important to determine whether an article spans just one page or several. We use a clustering system for identifying duplicate content on different URLs, as well as multi-page articles [23] with which we translate the corpus from labeled individual pages on the web to clusters of pages that represent articles. This is vital to avoid skewing volume-related metrics about longform. Finally, we are interested not only in questions about what longform looks like on the web but also how it compares with the general news publishing ecosystem, and thus include a large corpus of news articles as a baseline.

5.2 Extracted Data

To investigate longform along several dimensions, we extract many different signals from the data when developing our corpus. A number of these signals leverage internal systems that have previously generated the signal but for other purposes. For instance, we leverage systems that find the publication or byline date of each article (which, note, may differ dramatically from the first crawled date in instances where an archive has been brought online), as well as systems that determine the root site for a given url (which can be a difficult task given the number of possible domains and sub-domains under which a site could be hosted). A complete list of the signals collected is given in Table 3.

Table 3: Signals extracted for longform study.

Signal	Description
classifier score	score from the longform classifier
byline date	date the article was published
document date	date the article was first crawled
site name	name of the site the article is hosted on
anchor distribution	number of anchors observed and their observation dates

5.3 Results

With the corpus of longform documents in hand, we analyze numerous aspects of this corpus, performing an extensive study of longform on the web. Major aspects we consider are the current distribution of longform on the web (Section 5.3.1); the change to this distribution over time (Section 5.3.2); and characteristics unique to longform (Section 5.3.3). We obtain the following observations:

OBSERVATION 1. Distribution: *The distribution of sites publishing longform is heavy-tailed: longform can be found on many sites with a few high volume publishers. For some publishers, longform forms a significant fraction of published material. (Figure 4)*

OBSERVATION 2. Volume: *The sheer amount of longform content published each year is increasing though increasingly more slowly than the volume of news content published. Despite the decrease as a ratio, we observe an ever increasing amount of interest in longform when compared with news. (Figures 5b, 5b and 5c)*

OBSERVATION 3. Sites: *An increasing number of new sites are publishing longform. The increase in volume of longform publishing is due to many sites, not single large outliers. (Figures 6a and 6b)*

OBSERVATION 4. Interest: *There exists a quantifiable difference between how both publishers and readers interact with longform vs news content – readers interact with the content for longer, and publishers are more likely to bring older longform content online. (Figures 7 and 8)*

5.3.1 Distribution of Longform on the Web

In total, our corpus consists of more than 5.2 million documents currently available on the web and identified as longform, which we compare to a separate corpus of nearly 92 million news documents. To provide some initial insight into this corpus, in Figure 4 we chart the distribution of sites publishing longform. Not unexpectedly for web data, Figure 4a shows that this distribution is heavy-tailed: that is, there are many websites publishing just a few pieces of longform, and relatively few publishing many longform articles. The plot is shown in log-scale – the number of sites publishing beyond $e^7 \approx 1,000$ articles is hardly visible.

(a) Histogram of sites publishing different volumes of longform (log-scale).

(b) The percentage of longform found on high-volume sites (>1,000 articles).

Figure 4: The distribution of sites that publish longform is quite heavy-tailed (4a). Amongst high-publishing sites, several sites publish longform as a majority of their articles (4b) but many more include longform as a key component.

(a) Volume of longform produced every six months, for the past 10 years. (b) Ratio of the volume of longform produced vs. news for the past 10 years. (c) Average ratio of anchors in longform to news (a proxy for interest) over time.

Figure 5: Changes in the volume of and interest in longform produced over time. Although the number of longform articles produced yearly is increasing, the volume of longform produced relative to news is decreasing significantly. Despite this decrease, *interest* in longform with respect to news, measured via the ratio of average anchors, is steadily increasing.

A natural question about the few high-publishing sites ($>1,000$ articles per site) is whether they produce solely longform or if longform is a by-product of sites that publish a large amount of many article types. In Figure 4b, we see that amongst these sites, longform is a non-trivial amount of total production ($> 10\%$) for many of the sites, and in some cases, it is the majority of what is being produced. This indicates that longform plays an important role on these high-producing sites.

5.3.2 Changes in Longform over Time

Next, we investigate how longform has changed over time. We have both the byline date of each article and the first crawled date, a reasonable estimate for when the document was made available on the web. With these, we analyze temporal changes in longform. Our results here are biased by the existence of documents created in the past that have since been taken offline, a confounding factor that is unavoidable.

Some temporal trends are shown in Figure 5. In Figure 5a, we see that the volume of longform produced has increased significantly over the past 10 years, by a factor of more than 30 since 2005. However, as the web itself is growing, we must compare to other, similar sources. Figure 5b shows that the volume of longform published every six months has actually decreased significantly when compared to the volume of news articles produced in the same time period (by a factor of more than 7 since 2005). This decrease in volume relative to news may be evidence of cases where longform has been cut for other sources [51].

Despite the decrease in longform volume relative to news, Figure 5c shows that *interest* in longform compared to news has risen over the past 10 years, from 6 times to nearly 20 times as much. Interest in each of these groups was approximated via the total number of off-domain anchors, a classic measure of popularity used in e.g. PageRank [40]. All articles were grouped by publication date over 6 month periods, and the average number of off-domain anchors accrued (as of creation date) for articles in that period was calculated. Though older articles have had many years to accrue anchors, we find that average anchors for both longform and news are increasing, meaning that new content is popular for both. The fact that interest in longform is growing even more strongly provides evidence for a renaissance [5].

(a) The distribution of top sites over the past 10 years.

(b) The number of new longform sites added every 6 months.

Figure 6: The increase longform volume includes older sites producing more longform, as well as new sites producing longform for the first time.

Given the increase in longform volume, we aim to answer whether just a few sites are leading the movement, or if increases are apparent everywhere. Some answers to this are illustrated in Figure 6a, where we capture box plots of the volume of longform published at all sites over time. In particular, we aggregate the amount of longform published at each site during a six-month period, and then plot the distribution of these volumes, as in Figure 4a. This data is shown at six month intervals over a 10 year period. Given the heavy-tailed distribution of longform producing sites as described in 5.3.1, the box plots themselves are not visible.

However, it is interesting to see the outliers: there is not just a single site increasing in volume, but the number of total outliers and the degree of these outliers is increasing. This indicates that there are many, rather than just a few, sites that are driving the increase in longform volume.

We further investigate this in Figure 6b. This plot shows the number of new sites publishing longform, as indicated by the first longform byline date found on any article at that site. This indicates that it is not just that older sites are deciding to publish more longform, but rather new sites are either choosing to begin publishing longform (e.g. [buzzfeed.com](#) hiring a Longform Editor) or coming into existence for the purpose of publishing longform (e.g. [byliner.com](#) and other longform startups [36]).

5.3.3 Characteristics of Longform Interaction

Though we utilized several defining characteristics of longform while developing our classifier (Section 3), the corpus provides another opportunity to gather new insights about how different parties interact with longform versus news.

One such party is the publishers themselves. We measure their perception of the value of longform by measuring whether they bring older articles online, as this can be a laborious process. The “age” of an article at publication is calculated by taking the difference in the document crawl date and publication date.

Figure 7 plots the distribution of article ages for both longform and news. These numbers are scaled by the total number of longform or news articles, respectively, so that they can be fairly compared. Older longform articles at all ages are nearly twice as likely to be put online than news, including quite old articles, such as those published 15 years prior to coming online. This indicates that there may be more interest in these archived articles, either from the readers or publishers themselves. Given the low relative frequencies for both groups, however, it is also clear that the great majority of both longform and news articles on the web are put online directly after publication.

Figure 7: Distribution of the age of articles for longform and news. We see that longform articles are more likely to be put online after their by-line date than news articles.

In Figure 8, we investigate claims that longform is “ever-green”, or that it stays relevant longer than other journalistic styles [26]. We measure this by looking at the number of days until 2 months pass without our webcrawler finding any new anchors pointing to the document, which signifies that a systematic loss of interest has occurred. We show box-plots for these numbers over all news and longform articles. Though the results for both are quite heavy-tailed (most articles receive links for just a few days), longform articles tend to maintain external links, a proxy for interest, longer than

typical news articles. Some of the larger longform outliers retain anchors for nearly three months, which is significant given that most articles fall out of popular interest within a single day of being published.

Figure 8: The distribution of the number of days until a loss of interest (i.e. a break of 8 weeks of anchors) occurs in longform vs news. We see that longform tends to stay interesting for a longer period of time than news.

6. CONCLUSIONS

In this paper, we provide a way of selecting features to build a classifier that can successfully differentiate between longform and non-longform articles. Despite the huge variety in writing that is considered acceptable for a longform piece (as opposed to, for example, news reports), we show that using language features like parse structures, in combination with other simple document features, is surprisingly successful. We demonstrate how to generate viable training data and provide an experiment for labeling it with crowd-sourced raters. Our workflow in developing and using our classification system serves as an example for others hoping to successfully utilize the web for similar purposes.

Using the developed classifier, we build a corpus of longform and news documents from a web crawl and make several scientific discoveries about the longform ecosystem. Our findings include that the longform ecosystem is growing, both in terms of sheer number of documents published, as well as sites on the web creating this content, but that it is growing more slowly than other forms of publications, like online news reporting. While the average article published online does not receive sustained reader interest, we do find that successful longform articles live up to the ‘hype’ that they are more relevant and interesting for longer. These findings contribute to our general knowledge of what digital publishing looks like now and how it has changed over the past decade.

As to the future of longform, we believe that defining exactly what longform is will continue to grow more difficult, as it begins to include more immersive and multimedia content as all publishing moves online (see early examples like the New York Times’ “Snow Fall: The Avalanche at Tunnel Creek”). As a result of our study, the supply to the ecosystem appears to be quite healthy, though we reserve the right not to prognosticate about how this will change over the next decade. We leave the final word to Naomi Arnold that the future is “cautiously hopeful” [5].

Acknowledgements. The authors are grateful for help and advice from Pavan Desikan, Anand Shukla and Amar Subramanya, and to Dean Weesner for designing the rater experiment.

7. REFERENCES

- [1] The state of news media, *Pew Research Center*, 2013.
- [2] S. Abbar et al. Real-time recommendation of diverse related articles. In *WWW*, 2013.
- [3] A. Ahmed et al. Unified analysis of streaming news. In *WWW*, 2011.
- [4] S. Argamon, M. Koppel, and G. Avneri. Routing documents according to style. In *IIS Workshop*, 1998.
- [5] N. Arnold. The cautiously hopeful renaissance of longform journalism, 2014.
- [6] V. Ashok et al. Success with style: Using writing style to predict the success of novels. *Poetry*, 2013.
- [7] J. Bennet. Against ‘long-form’ journalism, *The Atlantic*, 2013.
- [8] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *WWW*, 2010.
- [9] A. Berger, S. Pietra, and V. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1996.
- [10] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.
- [11] T. Capote. The duke in his domain, *New Yorker*, 1957.
- [12] C. Cooper. The death of slow journalism, *American Journalism Review*, 2009.
- [13] C. Danescu-Niculescu-Mizil et al. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*, 2013.
- [14] M.-C. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC*, 2006.
- [15] L. DVorkin. Inside forbes: How longform journalism is finding its digital audience, *Forbes*, 2012.
- [16] A. Finn and N. Kushmerick. Learning to classify documents according to genre. *JASIST*, 2006.
- [17] A. Finn, N. Kushmerick, and B. Smyth. Genre classification and domain transfer for information filtering. In *Advances in Information Retrieval*. 2002.
- [18] K. Ganchev and M. Dredze. Small statistical models by random feature mixing. In *Workshop on Mobile NLP at ACL*, 2008.
- [19] M. Garber. Sit back, relax, and read that long story on your phone, *The Atlantic*, 2014.
- [20] M. Gaulon-Brain. Print media and television: Is longform bound for extinction?, *Ina Global*, 2013.
- [21] S. Gollapalli et al. Researcher homepage classification using unlabeled data. In *WWW*, 2013.
- [22] G. Greenwald and E. MacAskill. Nsa prism program taps in to user data of apple, google and others, *The Guardian*, 2013.
- [23] M. Henzinger. Finding near-duplicate web pages: A large-scale evaluation of algorithms. In *SIGIR*, 2006.
- [24] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. Ontonotes: the 90% solution. In *NAACL-HLT*, 2006.
- [25] C. Johnston. What buzzfeed’s evolution says about the future of longform journalism, *Poynter*, 2012.
- [26] Kaggle. Stumbleupon evergreen classification challenge, 2013.
<https://www.kaggle.com/c/stumbleupon>.
- [27] S. Kamdar. Highlighting content creators in search results. Inside Search, Google Search Blog, 2011.
- [28] S. Kandell. What i learned from a year of doing longform at buzzfeed, *The Big Round Table*, 2013.
- [29] H. Kwak et al. What is twitter, a social network or a news media? In *WWW*, 2010.
- [30] M. Lewis. Obama’s way, *Vanity Fair*, 2012.
- [31] J. Liu, P. Dolan, and E. Pedersen. Personalized news recommendation based on click behavior. In *IUI*, 2010.
- [32] A. Louis and A. Nenkova. What makes writing great? first experiments on article quality prediction in the science journalism domain. *TACL*, 1:341–352, 2013.
- [33] Y. Lv et al. Learning to model relatedness for news recommendation. In *WWW*, 2011.
- [34] J. Mahler. When ‘long-form’ is bad form, *The New York Times*, 2014.
- [35] F. Manjoo. You won’t finish this article, *Slate*, 2013.
- [36] K. McBride. Jill abramson startup to advance writers up to \$100k for longform work, *Poynter*, 2014.
- [37] I. Meuret. A short history of long-form journalism, *Ina Global*, 2013.
- [38] P. Nayak. In-depth articles in search results. Inside Search, Google Search Blog, 2013.
- [39] J. Nivre et al. The conll 2007 shared task on dependency parsing. In *EMNLP-CoNLL*, 2007.
- [40] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
- [41] S. Parker. Buzzfeed’s success does not mean we should be slaves to clicks, *The Guardian*, 2014.
- [42] N. Paumgarten. Up and then down, *New Yorker*, 2008.
- [43] C. Plante. Street fighter: The movie - what went wrong, *Polygon Magazine*, 2014.
- [44] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*. 2002.
- [45] R. Rieder. Long-form journalism makes a comeback, *USA Today*, 2013.
- [46] Salmon. Jeff bezos and his journalists. *Reuters*, 2013.
- [47] M. Santini et al. Implementing a characterization of genre for automatic genre identification of web pages. In *COLING/ACL*, 2006.
- [48] J. Sappell and R. W. Welkos. The scientology story. *The Los Angeles Times*, 1990.
- [49] S. Sharoff. Classifying web corpora into domain and genre using automatic feature identification. In *Web as Corpus Workshop*, 2007.
- [50] B. Smith. What the longform backlash is all about, *Medium.com*, 2014.
- [51] D. Starkman. Major papers’ longform meltdown, *Columbia Journalism Review*, 2013.
- [52] G. Talese. Frank sinatra has a cold, *Esquire*, 1966.
- [53] A. Tumasjan et al. Predicting elections with twitter. In *ICWSM*, 2010.
- [54] D. F. Wallace. The string theory, *Esquire*, 1996.
- [55] H. Wang et al. Joint relevance and freshness learning from clickthroughs for news search. In *WWW*, 2012.
- [56] G. Wong and W. Mason. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 1985.