BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos

Giovanni C. Santia, Jake Ryland Williams
Department of Information Science,
College of Computing and Informatics,
Drexel University, 30 North 33rd Street,
Philadelphia, Pennsylvania 19104,
{gs495,jw3477}@drexel.edu

Abstract

Veracity assessment of news and social bot detection have become two of the most pressing issues for social media platforms, yet current gold-standard data are limited. This paper presents a leap forward in the development of a sizeable and feature rich gold-standard dataset. The dataset was built by using a collection of news items posted to Facebook by nine news outlets during September 2016, which were annotated for veracity by BuzzFeed. These articles were refined beyond binary annotation to the four categories: *mostly true*, *mostly false*, *mixture of true and false*, and *no factual content*. Our contribution integrates data on Facebook comments and reactions publicly available on the platform's *Graph API*, and provides tailored tools for accessing news article web content. The features of the accessed articles include body text, images, links, Facebook plugin comments, Disqus plugin comments, and embedded tweets. Embedded tweets provide a potent possible avenue for expansion across social media platforms. Upon development, this utility yielded over 1.6 million text items, making it over 400 times larger than the current gold-standard. The resulting dataset—*BuzzFace*—is presently the most extensive created, and allows for more robust machine learning applications to news veracity assessment and social bot detection than ever before.

Introduction

As the internet becomes an ever-increasing presence in the life of the average person, more and more obtain their news from Facebook and other forms of social media (Gottfried and Shearer 2016). Since this dissemination of news content is by and large unsupervised and often strictly user-generated, quality control has become a pressing concern. Clearly, misinformation on the internet is not a new problem, as fact-checking websites such as Snopes have existed since at least 1994. What is new is this meteoric rise in social media which has made it easier than ever before for organizations to produce and spread news content of questionable validity to massive audiences (Chen, Conroy, and Rubin 2015). The spread of intentional misinformation through online forums during the 2016 Brexit vote and U.S. Presidential election (Howard and Kollanyi 2016; Howard, Kollanyi, and Woolley 2016) have put the spotlight on what has recently been dubbed "fake news". Facebook and other social media corporations have made attempts to counter this manufacture of misleading news content, but it has only become more prominent (Weedon, Nuland, and Stamos 2017).

Since shutting down the production of all outlets that produce such content would be nearly impossible, the main method being explored for the systematic squelching of this content is detection. An algorithm that would take a news article and its associated features and assign a veracity score would prove a potent weapon in combating misinformation online. Unfortunately, little progress has been made on such an algorithm (Conroy, Rubin, and Chen 2015). This holds true for a multitude of reasons, perhaps the most important being the lack of gold-standard data on which to train models (Rubin, Chen, and Conroy 2015). This problem is largely a matter of scope; the validity analysis of the content of thousands of news articles of non-trivial veracity requires significant vetting and input of time. Recently, a BuzzFeed News investigation has rendered one such dataset, whose potential we highlight.

An additional area of concern with respect to the propagation of information on social media platforms is social bots, which are often the means by which questionable news content is spread (Ferrara et al. 2016). Social bots are automated users of social media platforms which promote specific ideologies. It is widely thought that the misinformation campaigns associated with the 2016 U.S. Presidential Election and the Brexit vote were enacted by large numbers of coordinated social bots (Howard and Kollanyi 2016; Howard, Kollanyi, and Woolley 2016). Unfortunately, one of the largest factors contributing to the rise of social bots in everyday online discourse is the difficulty in their detection (Ferrara et al. 2016). Due to this, many experts are not even in agreement as to the actual scope of the problem, but a common figure cited is that between 9% and 15% of active users on Twitter are automated (Varol et al. 2017). This is a huge proportion of users, and clearly has already had massive impacts on not only social media ecosystems, but society at large. Any progress made towards the creation of algorithms for detection of these social bots would have massive implications in many fields, but unfortunately similarly to the problem of misleading news content detection, there is an alarming shortage of gold-standard data. The only

data sets of any relevance that we have been able to find have dealt with Twitter. The data set that we detail here — *BuzzFace* — will be particularly interesting to those that wish to study the social bots of Facebook.

The initial dataset created by BuzzFeed was formed using a sample of news articles posted to Facebook by a select group of news outlets during a specific time period (Silverman et al. 2016). Each article was read and analyzed for veracity, and given a categorization. While it is in and of itself a useful dataset, we have identified this opportunity for its enrichment via the features that come along with social media posts. As the articles were all posted to Facebook, the Facebook Graph API allows for the collection of data on "reactions", Facebook comments, and various metadata. Additionally, since many of these Facebook posts link to outside web pages, content acquisition can be performed to obtain the article text along with images, links, and embedded tweets associated with each article. Finally, many news outlets allow for additional commentary on the actual article web pages themselves—using either or both of the Facebook Comments and Disqus Comments System plugins. This allows for up to two additional sources of commentary associated with each article.

While websites and platforms such as Facebook offer tremendously valuable data publicly and through their APIs, these services are ephemerally dependent on variations in terms of service and considerations for user privacy. However, these issues are as much subject to current events as they are to user preference settings and the whims of administrators. Since *BuzzFace*'s inception ($\approx$ 1 year), Facebook has updated its API from version 2.8 to 2.12, which has included the removal of user-level data of reactions from public access. In addition to this, many user comments, news articles, and even an entire news outlet have been deleted or hidden—Freedom Daily's page is no longer accessable, making neither its posts nor its comments available in this final release. Additionally, while the dataset's initial access resulted in user-identifying information as a component of the Facebook comments, this information is now restricted. Thus, present retrieval from the Graph API renders commentary that is lacking this information, potentially hobbling the dataset from user-based analyses. However, seeing the value of these data we address these issues here by independently releasing ego identifiers. Ultimately, these issues highlight the ephemerality and evolution of social media, and for the data on egos we release, make the *BuzzFace* dataset all the more valuable. *BuzzFace* is hosted at (Santia and Williams).

Related work

For such an important contemporary issue, there have been relatively few scholarly studies produced with the intention to aid the assessment of news veracity online, particularly in the context of social media. Indeed, we have found no such datasets which pertain to Facebook. Clearly, this is not a new problem, as fact-checking websites such as Snopes have existed since at least 1994. Considering that Facebook is by far the largest social media platform and the source of many Americans' daily news (Center 2015), this fact came as a surprise. Perhaps this is due to the difficulty researchers face in obtaining large gold-standard datasets using Facebook's Graph API; Facebook is notoriously protective of their data. This lack of Facebook news veracity data has led to a severe shortage of gold-standard datasets for potential investigators to work with. This shortage has led to extreme difficulty in the creation of reliable news veracity classification algorithms (Rubin, Chen, and Conroy 2015).

The current gold-standard in social media veracity assessment is the corpus which was used in the shared task at *SemEval-2017*, titled "RumourEval: Determining rumour veracity and support for rumours" (Derczynski et al. 2017). This task required participants to determine the veracity of a given set of tweets. Along with each of these tweets was provided the associated "conversation" of tweets stemming from the parent. This grouping of tweets naturally generates a tree structure. The participants of the task set out to use the thread associated with a parent tweet to determine its veracity, with the only available veracity classifications being *true* or *false*. Such a classification leaves very little room for ambiguity and does not provide the user with tools for a descriptive annotation. Oftentimes, misleading news items are so effective precisely because they contain just enough true information so as to come off as legitimate, and then the falsehoods therein are all the more effective at misleading the reader (Ecker et al. 2014). This dataset contains 297 threads, each with a unique parent tweet. The reply tweets total up to $4,222$, which makes for a dataset of $4,519$ tweets. Relatively speaking, this is quite a small amount of data to work with. In addition, the data allows the user to train an algorithm to determine the veracity of a single tweet, which at the time could be no longer than 140 characters. Naturally, the content of a tweet is miniscule in comparison to the content of the standard news article. This brevity yields a much less complicated text object to analyze. Clearly it is easier to inject a variety of misleading and true statements into documents of greater length. Larger datasets which also focus on small pieces of text exist — such as LIAR at $12,386$ items (Wang 2017) — but they lack the user-generated content associated with social media. Similar to LIAR is the dataset detailed in (Vlachos and Riedel 2014) but this is quite a bit smaller at 106 short text items.

Another dataset of interest is the "Fake News Challenge" (Team 2018). The current goal of the dataset is to facilitate production of algorithms which classify the stance of the body of an article relative to the claim made in the title. The challenge asks users to make this classification into one of four categories: *agrees*, *disagrees*, *discusses*, and *unrelated*. The presence here of more than two categorizations is certainly an advantage that this dataset has over the previously discussed RumourEval dataset. Unfortunately, this dataset is designed only to facilitate this process — also called stance detection, which is not equivalent to news veracity assessment. The authors claim that automating stance detection is an important first step in the creation of veracity algorithms. While this may be the case, in its current form, this dataset provides no help in researchers attempting to create better mechanisms for veracity assessment. The data itself consists of the body text of $2,532$ articles coupled with $49,972$ titles

which are each assigned a corresponding body text. Clearly, many of the body texts will have several different titles assigned to them. The participants are also given one of the previously-described labels for each of the titles, in order to facilitate training. There is no indication whatsoever in the dataset as to the veracity of the articles provided. Still, this is a reasonably-sized dataset which perhaps in the future will show much merit in veracity assessment. It is interesting to note that the authors state this dataset was derived from the Emergent online news veracity classifier, which was created by Craig Silverman, who was the main leader of the BuzzFeed dataset we have based our work on (Silverman et al. 2016).

Content

BuzzFeed Dataset

The dataset provided by BuzzFeed consists of $2,282$ news articles, along with several Facebook features (number of likes, etc.) and the assigned veracity rating. The articles include all posts from seven weekdays in September 2016 made through the following nine Facebook news pages: *ABC News Politics*, *Addicting Info*, *CNN Politics*, *Eagle Rising*, *Freedom Daily*, *Occupy Democrats*, *Politico*, *Right Wing News*, and *The Other 98%*. This time frame—the height of the 2016 Presidential Election—saw increased public awareness of the online information veracity issue. The outlets were chosen such that they represented various possible political biases: mainstream, left-leaning, and right-leaning. The mainstream outlets were *ABC News Politics*, *CNN Politics*, and *Politico*. The left-leaning outlets were *Addicting Info*, *Occupy Democrats*, and *The Other 98%*. The remaining three outlets were right-leaning. The BuzzFeed report (Silverman et al. 2016) exhibited the timely nature of the problem, with the more-partisan outlets publishing false and misleading information more than 20% of the time.

While it may seem natural to simply use binary categories when assigning the news items veracity labels (namely, *true* and *false*), the curators of the BuzzFeed dataset decided to take a more nuanced approach and used the following four: *mostly true*, *mostly false*, *mixture of true and false*, and *no factual content*. *Mostly true* and *mostly false* are straightforward and used when the majority of the information in the news item is either accurate or inaccurate, respectively. *Mixture of true and false* is chosen when the inaccurate information is roughly equal to the accurate, or when the news item is based on unconfirmed information. Finally, *no factual content* is used in the case of posts which are opinion, comics, satire, or other posts that do not make a factual claim (Silverman et al. 2016). This system is more informative than a simple truth dichotomy as it recognizes the significant volume of content online that simply contains no factual information. Moreover, this categorization will allow researchers to study perceived credibility when truth and falsehoods are mixed.

Many of the other features provided in the dataset are made obsolete by the additional processing. An essential feature for each article is the Facebook ID, which allows for easy use of the Facebook Graph API. At the initial time of access, this provided use of the API for data on $2,263$ of the articles (fewer than 1% were deleted or had no comments). Another useful feature is Post Type, which categorizes each article as either a *link*, *photo*, or *video*. This is crucial information for the content acquisition process, as there is generally no text to access in a photo or video. Other useful features provided include the counts of the numbers of *shares*, *reactions*, and *comments* on Facebook. However, these numbers were tabulated in 2016 and since the production of this data, have changed. At the time of initial access, the Facebook Graph API allowed for retrieval of not only the correct counts, but the data objects representing the *shares*, *reactions*, and *comments* themselves in an ongoing fashion. However, since the Graph API transitioned to version 2.12 on January 30$^{\text{th}}$, Facebook ceased to make user-level reactions available. Thus, user-level reactions are no longer an accessible portion of BuzzFace.

Processed Data

A description of our contribution to BuzzFace follows. While individuals are permitted to perform the API calls and access web content in production of the discussed data, the different components fall under a variety of licensing agreements that prevent their full, collated publication. Thus, we provide only the Python scripts necessary to populate the data.

Facebook Comments and Reactions The content of the Facebook posts are thoroughly enriched on the platform by active reader commentary. We have used the Facebook Graph API to collect the comments associated with each article to create a dataset of over 1.6 million comments discussing the news content. The only similarly-focused social-information veracity assessment resource (Derczynski et al. 2017) covers approximately 300 claims and $4,000$ follow-up replies. Thus, BuzzFace covers approximately 7 times the number of stories, and over 400 times the number of individual messages than the state-of-the-art.

This final result of over 1.6 million comments is quite a bit higher than the total summation of comments on these articles as reported by BuzzFeed themselves, which was $1,176,713$ comments. While part of this may be contributed to the passage of time which yields additional comments, this cannot explain the massive gap. It turns out that BuzzFeed mistakenly under-reported the number of comments on each article when they published their original dataset, which only makes analysis of it all the more valuable. It turns out that there are two distinct types of Facebook comments, which we deem "top-level" and "replies". Top-level comments are those which are made directly in response to the Facebook Object in question, while replies are comments made in response to a particular top-level comment. The first comment left on a Facebook Object (a wide-range of Facebook items including posts made by individual users and pages) must be a top-level comment, as at that time there are no top-level comments to reply to. After this first comment is made, any user leaving additional commentary now has the choice to respond to the original post itself or to a top-level comment. This is as far as the nested structure of the

Veracity category	# Articles	# Top-level comments	# Replies	Total # comments	Comment rate	Avg. characters per comment
No factual content	259	564,086	132,029	696,115	2,687.70	81.60
Mixture t/f	244	188,184	41,988	230,172	943.33	109.08
Mostly false	104	49,624	9,454	59,078	568.06	97.20
Mostly true	1,656	516,153	182,575	698,728	421.94	155.97
All	2,263	1,318,047	366,046	1,684,093	744.19	108.76

Table 1: Decomposition of article veracity by user comments. Nearly three quarters of articles (73.18%) consisted of *mostly-true* content, whereas fewer than half (41.49%) of all comments focused on these. *Mostly-true* factual content stands out well below the other veracity categories in the number of comments per article, while articles with *no factual content* exhibited extremely high activity. Independent of this, *mostly-true* factual content is also strongly marked by much longer comments, which average from more than 50% to almost twice the size of those from other categories.

comment threads goes; when one leaves a reply to one of the "replies", it is itself again amongst the "replies". Thus in the same way that each post has a commentary thread consisting of top-level comments, one may consider each top-level comment as its own "post" with its own thread of replies. For reasons unknown to us, when the Facebook Graph API is used to obtain "all comments" on an Object, only a list of the top-level comments is returned. Therein lies the mistake BuzzFeed most surely made; they reasonably figured that to obtain all comments on the posts it would be adequate to merely use the API method which gives "all" the comments. It turns out that in order to obtain all of the replies as well, it is necessary to call the get comments API command on each of the over 1.3 million top-level comments in turn and append the results to the dataset. This process gave us an additional $366,046$ comments. We made sure to preserve this threaded-structure of the commentary, and our methods for doing so are discussed in a later section.

Our original access to the Graph API (versions 2.8–2.11) rendered comments with all user IDs and names as available fields (keyed by "from" in the JSON response comment objects). Thus, the original integration rendered a version of *BuzzFace* that might be studied for user-level interactions and user groupings of comments. However, with the release of version 2.12, the Graph API's documentation stated:

> On February 5th, 2018, User information will not be included in responses unless you make the request with a Page access token. This only applies to Comments on Pages and Posts on Pages.

which indicated that only page owners would receive user identifying information, going forward. Thus, we can only infer that researchers who wish to access *BuzzFace* will not be provided with user information from the Graph API. While we cannot release this information, we maintain the utility of *BuzzFace*'s user-level information by releasing anonymized, ego identifiers (Ego-IDs) associated to the Graph API's comment IDs (see Sec. Structure for more details).

Plugin comments In addition to the comments made on each of the Facebook posts, many of the articles possess a comments section on the outlet's website itself. Every outlet that allows for such comments employed the Facebook Comments plugin and/or the Disqus Comment System (*Eagle Rising* has a separate comments section for each). Table 2 explores the distribution of these plugin comments by outlet. Obtaining the Facebook plugin comments was a similar process to obtaining the comments on Facebook itself, with the additional step of needing to query the Graph API for the IDs of the article (here they are not simply in the URL). This process yielded an additional $82,090$ Facebook comments. It is important to note these comments are produced by users who may never have accessed the Facebook posts annotated by BuzzFeed and could simply be avid followers of the outlets in question. These comments may be tapping into an entirely separate demographic to the other set of Facebook comments. This is almost certainly the case for the Disqus plugin comments.

The Disqus Comment System is one of the most popular commenting systems employed on the internet. The demographic of users generating Disqus comments is vastly different from those creating the Facebook comments. This is because the Disqus platform not only allows users to sign in using their pre-existing Google, Twitter, or Facebook accounts, but users may create a custom Disqus profile. This gives users without social media accounts a chance to create comments. The process of obtaining these comments was fairly similar to that of Facebook; first, the proper ID had to be extracted (this was an outlet-dependent endeavor), then the correct calls to the Disqus API were made. The structure of Disqus threads is overall very similar to those of Facebook. Users may leave new comments or submit replies to, share, and "like" existent comments. The structure of this data is also very similar to the data obtained from the Facebook Graph API.

Disparity It is clear that the number of Facebook comments made on the actual Facebook posts themselves dwarfs the plugin comments. It would be informative to launch a study into why. A possible explanation is that a sizeable number of the Facebook users which consumed the content of the outlets merely read the title of the news articles without clicking through to their texts before commenting. In any case, these plugin comments are still a valuable contribution to the dataset for their differing demographic.

Outlet	% articles deleted	% third party	# Facebook comments	Facebook comment rate	# Disqus comments	Disqus comment rate	# tweets
ABC News Politics	12.04	3.704	-	-	28,877	277.66	45
Addicting Info	-	27.61	-	-	-	-	77
CNN Politics	6.36	2.12	-	-	-	-	97
Eagle Rising	0.866	56.28	11,414	113.01	2,054	20.34	7
Freedom Daily	8.257	21.10	2,266	26.35	-	-	9
Occupy Democrats	-	14.89	-	-	-	-	-
Politico	0.432	0.864	66,875	145.70	-	-	10
Right Wing News	8.527	37.21	1,535	9.48	-	-	26
The Other 98%	5.882	94.12	-	-	-	-	-
All	4.049	20.42	82,090	101.60	30,931	150.88	271

Table 2: Article deletion, third-party status, Facebook plugin comments Disqus plugin comments, and tweets by outlet at the time of initial access. Note: since the time of initial access the Freedom Daily page ceased to be publicly available.

Other Social Media

The BuzzFeed dataset strictly dealt with news items posted to Facebook. While Facebook makes up a sizeable portion of the social media sphere, it is by no means comprehensive (Gottfried and Shearer 2016). Any expansion of the dataset to other platforms would yield massive amounts of new robust data to explore. As it stands, our dataset is well-positioned to incorporate both Twitter and Reddit.

Twitter The news items included in the BuzzFeed dataset were all created in September 2016 and thus are almost entirely focused on the United States Presidential Election. As both major campaigns were very active on Twitter, many of the articles made references to tweets. Twitter provides the tools necessary to web developers to create embedded tweets in their web pages, and this provides an easy mechanism to link our data with Twitter. When performing our web content acquisition on the articles we made sure to find all such instances of embedded tweets and record the URLs of the tweets they linked to. Table 2 presents more information on their occurences. These harvested tweets would yield a significant amount of additional data to analyze using the Twitter API. For example, the number of favorites and retweets along with all replies to the tweets would be simple to obtain and informative.

Reddit An additional social media platform that provides many users with their daily news is Reddit. In particular, there are several sections of Reddit—"subreddits"—where users may only post links to news articles. These function in a very similar way to the Facebook news posts in the BuzzFeed data: users may "like" and comment. We could use the Reddit API to search these subreddits for any of the articles included in the BuzzFeed dataset and extend our dataset with the corresponding Reddit comments and other pertinent features.

Quality

The original annotations which we have based our dataset on were completed by a team of journalists at BuzzFeed (Silverman et al. 2016). This team included several journalists whose careers rely heavily on the ability to verify or reject news reports. Thus we are treating these annotations as gold-standard data. The BuzzFeed team made sure to keep the data largely representative of multiple types of news outlets by selecting them from the mainstream, left-leaning, and right-leaning categories. It is important to note that each of the outlets that the team chose had been "verified" by Facebook, and thus in an indirect fashion have been deemed as being more credible than other, non-verified, news outlets on the platform. In addition, the team also detailed (Silverman et al. 2016) the fact that they not only were checking for the accuracy of the information in the text of the articles, but would also label the articles as a *mixture of true and false* if the content of the article was true for the most part, but did not match the claims made in the title or caption. They found that oftentimes Facebook share lines and/or titles would inject misinformation or misleading information into an otherwise respectable article. The team even ended up changing some of the annotations for the articles after receiving feedback which proved they had made the wrong choices. These dimensions of their analysis, along with their qualifications and resolve to find the truth have made for a compelling dataset. Indeed, studies have already been completed using this BuzzFeed data set as a starting point. The results of (Potthast et al. 2017) in particular are of interest. They focus on a stylometric analysis of the body text of the articles themselves, forgoing the use of Facebook to provide extra content.

Enrichment by additional data makes for a more robust dataset. While the BuzzFeed dataset is useful and a welcome record of veracity, it alone does not possess language to process, limiting its machine learning development capacity. Alongside the posts, our addition of the news article content and Facebook and plugin comments provides a sizeable collection of text and multimedia that could be used for multiple learning tasks. We have managed to capture 2263 of the posts and their commentary (99.17%), which all came from the Facebook Graph API in pristine condition. The data was then organized and munged by our scripts to allow for

Veracity category	# Articles	# Top-level comments	# Replies	Total # comments	Comment rate	Avg. characters per comment
ABC News Politics (784,622 fans)						
No factual content	26	2,361	1,556	3,917	150.65	121.03
Mixture t/f	2	93	31	124	62	142.12
Mostly false	0	0	0	0	-	-
Mostly true	172	9,976	5,769	15,745	91.54	144.62
All	200	12,430	7,356	19,786	98.93	139.94
Addicting Info (1,427,134 fans)						
No factual content	11	2,088	1,266	3,354	304.91	104.18
Mixture t/f	25	9,955	3,735	13,690	547.6	99.43
Mostly false	8	4,954	1,074	6,028	753.5	94.12
Mostly true	96	30,804	12,873	43,677	454.97	94.46
All	129	47,801	18,948	66,749	517.43	95.94
CNN Politics (2,681,981 fans)						
No factual content	20	8,189	3,589	11,778	588.9	200.96
Mixture t/f	4	936	547	1,483	370.75	234.54
Mostly false	0	0	0	0	-	-
Mostly true	385	108,852	66,020	174,845	454.14	227.60
All	389	117,977	70,129	188,106	483.56	225.99
Eagle Rising (689,483 fans)						
No factual content	81	3,569	396	3,965	48.95	84.74
Mixture t/f	54	6,569	779	7,348	136.07	89.91
Mostly false	30	2,316	514	2,830	94.33	126.74
Mostly true	121	6,731	768	7,499	61.98	92.62
All	205	19,185	2,457	21,642	105.57	94.72
Freedom Daily (2,658,870 fans)						
No factual content	4	1,277	191	1,468	367	101.84
Mixture t/f	26	12,066	1,146	13,212	508.15	87.49
Mostly false	26	12,599	1,818	14,417	554.5	100.65
Mostly true	56	21,076	2,767	23,843	425.77	91.88
All	112	47,018	5,922	52,940	472.68	93.45
Occupy Democrats (7,111,843 fans)						
No factual content	65	433,256	163,455	596,711	9,180.17	74.49
Mixture t/f	33	102,710	33,115	135,825	4,115.91	111.41
Mostly false	9	11,987	4,278	16,265	1,807.22	102.66
Mostly true	102	159,213	70,937	230,150	2,256.37	127.52
All	209	707,166	271,785	978,951	4,683.98	92.55
Politico (1,762,151 fans)						
No factual content	6	267	196	463	77.17	335.49
Mixture t/f	2	2,517	1,370	3,887	1,943.5	117.22
Mostly false	0	0	0	0	-	-
Mostly true	528	69,174	39,579	108,753	205.97	192.62
All	534	71,958	41,145	113,103	211.80	190.61
Right Wing News (3,561,400 fans)						
No factual content	11	3,770	628	4,398	399.82	81.12
Mixture t/f	89	38,343	3,806	42,149	473.58	94.10
Mostly false	26	10,872	1,720	12,592	484.31	78.54
Mostly true	142	39,489	4,446	43,935	309.40	86.60
All	268	92,474	10,600	103,074	384.60	88.45
The Other 98% (5,520,002 fans)						
No factual content	40	45,330	24,919	70,249	1,756.23	113.67
Mixture t/f	10	11,103	6,380	17,483	1,748.3	139.04
Mostly false	5	5,022	2,680	7,702	1,540.4	108.75
Mostly true	67	49,635	28,213	77,848	1,161.91	123.38
All	122	111,090	62,192	173,282	1,420.34	120.37

Table 3: Statistics detailing Facebook comments and other items by outlet. The number of Facebook fans that each of the outlets currently has as of time of writing is provided next to the outlet name. The distribution of the articles studied by outlet is also provided. The left-leaning outlets have by far the highest comment rates, while the mainstream outlets have the highest average comment lengths. These are both possible indicators of social bot activity. Note: since the time of initial access the Freedom Daily page ceased to be publicly available.

quick analysis and easy-access. The storage of the vast quantities of data in JSON objects allows for quick retrieval of desired data subsets and an intuitive and descriptive means of organization of features. The JSON objects representing the comments themselves are automatically chronologically sorted to maximize simplicity. A key advantage of our con-

tribution to the dataset is the fact that the majority of it is non-static, since commentary continues to accrue over time. Users are continually adding new comments long after the initial post dates of the news items and reacting to said commentary. Since the BuzzFeed dataset was harvested, the total number of comments on the news items has increased by over $50,000$. Due to the fact that we are distributing scripts for the creation of local datasets for the users, this allows for the data to continually grow in size in this fashion, and is not limited to just that which we discuss here.

Structure

As stated previously, the actual files provided will merely be a suite of Python scripts which will perform all of the necessary web content acquisition, API requests, creation of directory hierarchies on disk, and writing of the data. Once this process is completed, the user will be left with the entire dataset at their disposal, along with a custom-made API that allows for efficient slicing of the data.

Data

First the data corresponding to each of the news items is collected and saved, and then aggregated into unifying data structures to enable quick retrieval. The main directory will have 9 sub-directories (one for each of the news outlets) which contain directories for each of the $2,282$ news items annotated by BuzzFeed. Each directory will possess the associated Facebook post ID as its name and contain the following JSON files: *attach.json*, *comments.json*, *posts.json*, *replies.json*, and *scraped.json*.

1. *attach.json* includes information on the attachments to the post, including the images, videos, links, title, and subtitles. Keys which provide the URLs of all of these features are also present.
2. *comments.json* is a list of data pertaining to all the top-level comments made on the post. The precise features of comment objects that populate this file are identical to those provided by Facebook's Graph API.
3. *posts.json* details post metadata, including: caption, creation time, its Facebook post ID, its link, the message, the name, pictures, number of shares, the type (link, image, or video), and the last time it was edited.
4. *replies.json* is again a list of the comments made on the post, but this time including both the top-level and replies. The file is formatted to represent the threaded structure of Facebook commentary. Each top-level comment is represented as a JSON object with the following keys:
   (a) *created_time* simply yields the time the comment was made.
   (b) *id* is the Facebook comment ID associated with the comment.
   (c) *message* is the text of the comment.
   (d) *replies* is a list of JSON objects representing all the replies made to this top-level comment. Replies have the same structure as the top-level comments, except they are missing the *replies* key. If this list is empty, it means that no replies were made.
5. *scraped.json* is the result of the web content acquisition applied to the news items which linked to actual text articles on other webpages. This is a JSON object with the following keys:
   (a) *links* which is a list of all the links contained within the body of the article, along with text.
   (b) *pictures* is a list of the URLs of all the pictures in the body of the article, along with their captions.
   (c) *body* is simply the text of the body of the article.
   (d) *tweets* is a list of all the embedded tweets in the body of the article.
   (e) *comments* is a list of all the comments made on the article using the Facebook Comments Plugin. These are structured just like actual comments made on the Facebook platform themselves, without the replies key, as they are all top-level.
   (f) *DisqComm* is a list of all the comments made on the article using the Disqus Comments Plugin. These are again represented as JSON objects, but they contain so many keys and values that it would be quite lengthy to describe them all. The Disqus API provides much more information than the Facebook Graph API.

API

As stated above, the full dataset is massive with a multitude of different types of data. In order to facilitate analysis of this data, we have created an effective API to allow the user to extract specific subsets. This API is provided in the form of multiple Python scripts which are well-documented. In order to initiate the API, the user must simply run the *main.py* file included with the data. The API has methods which allow for the analysis of the commentary either by user or by thread. Since in both the data and the Facebook Graph API there is a clear distinction made between the top-level comments and the replies, we have included the ability for the user to specify which type of comments they would like to analyze when using the majority of the API methods: all comments, just top-level, or just replies. These are the methods available to the user (all of them allow for the choice of comment level except for *cutThread*):

- *Text* - there are versions of this function for both a single User or a Thread. It simply returns a list of all the text of the comments in question.
- *Times* - again there are versions of this function for User or Thread. It returns a list of all the times the comments in question were made, in datetime.datetime format.
- *TextTimes* - includes versions for User or Thread. This function returns the output of the previous two functions zipped into a single list of tuples.
- *Response* - includes version for User or Thread. Returns a list of the response times of the comments in question as a number of seconds. We define response times to be the time that passed between the comment and the previous top-level comment for top-level comments, and the time between the comment and the previous reply for replies. In the case that the comment is the first top-level comment

in a thread, the response time is simply the time passed between the original post time of the thread and the comment. When the comment is the first reply to a top-level comment, the response time is the time that passed between the comment and its top-level parent comment.

- *ThreadCounter* - this is a method only for the User class. It will return a Counter showing the Facebook thread IDs that the user commented in along with their frequencies.
- *UserCounter* - this is a method only for the Thread class. It will return a Counter showing the Ego-IDs for the users that added comments to the thread and their frequencies.
- *CutThread* - this is a method only for the Thread class. It allows the user to analyze a given thread only up until a specified time. This shortened thread may then be used in the same way as a complete thread.

Potential uses

When performing a literature review in regards to similar datasets and their applications, we found nothing similar to *BuzzFace*. Pristine Facebook data is notoriously difficult to obtain (Rieder 2013), and thus it made sense that we found little to no large datasets which focused on veracity assessment that incorporated it. Not only this, but we also found no such datasets which focused on social bot detection on the platform. It is important to note there was a selection of studies completed which sought social bot detection techniques on other social media, particularly Twitter. Both news veracity assessment and social bot detection have become incredibly important and popular areas of focus in Natural Language Processing and Computer Science research in recent times due to high profile and large-scale political events around the globe. An intriguing potential use for the dataset we present here is to create and train machine learning models for these two avenues on the Facebook platform.

News veracity assessment

To date, much of the attempts at classification of news articles into categories of veracity have relied solely on the content of the articles, and has not paid due attention to associated user-generated content. Our addition of the massive quantities of text coupled with each news item will allow for researchers to have far more data to work with when creating their models than before, and this may lead to more reliable and effective veracity assessment. While the comments themselves do not come paired with their own veracity annotations (there are simply too many of them for a small team to have annotated them by hand, and additionally oftentimes Facebook comments are difficult to classify as true or false as they are simply an expression of opinion), each comment is paired with the veracity annotation of its parent post. Thus a potential investigator may be able to find features of comments which most likely indicate that the comment was made on a post of questionable validity, and then use this information to take the comments associated with a novel news item and make a veracity classification. Such a system could thus be used to make such classifications in real-time shortly after the items are posted, needing access only to the article and its commentary.

The *BuzzFace* data has important characteristics that indicate its quality for such development of a machine learning model for news veracity assessment. Breaking down the user comments by veracity, we see that on average, articles labeled as *mostly true* received comments at a sizeably diminished rate (421.94 comments per article) of those labeled as *mostly false* (568.06 comments per article). Strikingly, articles with *no factual content* exhibited extremely high comment rates ($2,687.70$ comments per article). Additionally, we note that comments on articles labeled as *mostly true* were approximately twice as long (155.97 characters per comment), on average, as those of their *mostly false* counterparts (97.20 characters per comment), with articles of *no factual content* once again at the extreme opposite end of the spectrum (81.60 characters per comment). These variations (few, but long comments) are present only for the articles labeled as *mostly true*—as evidenced by Table 1—a clear signature for true factual content. Finally, a finer-grained breakdown of BuzzFace by outlet is provided in Table 3, where it can be seen that Occupy Democrats articles astonishingly received more than half of all comments.

The presence of behavioral differences that may be leveraged in veracity assessment are highlighted by these findings. These signatures do not exist in the other state-of-the-art dataset (Derczynski et al. 2017), where rumors labeled as true neither received more nor longer replies (however, that source was Twitter, having a short-form, character limit of 140 at the time of completion). Moreover, the shared task associated with these data resulted in none of the 13 submitted systems outperforming a baseline of random assessment by the rate of false rumors (Derczynski et al. 2017). This suggests the existing resources may lack sufficient size and/or quality to advance system development.

Social bot detection

Social bots have increasingly made their presence known to Facebook users in recent years (Ferrara et al. 2016). Unfortunately, since there was not much awareness of this issue until now, there has been little scholarly work done on their detection. Apart from the issue of the fast-changing pace of the social media landscape, an additional factor which may contribute to this lack of progress is the previously-mentioned difficulty in acquisition of Facebook data. As many Facebook users post quite a bit of personal information, Facebook is less willing to provide its data to the public than other—more anonymous—social media platforms such as Twitter. While this is reassuring to the average user, it could potentially make Facebook a trivial platform for a nefarious actor to infest with social bots. Considering that our dataset comprises news articles posted during the peak of activity during the 2016 U.S. Presidential Election, which was one of the main events which brought social bots to the attention of the public and the world at large, it is almost certain that we have captured social bot activity. Researchers interested in trying to trace social bots and their impact on the Election would find much of interest in the data we present. More generally, the data could be used to create new systems for making classifications of Facebook users as either humans or social bots. While the dataset does not contain

annotations labeling the $843,690$ users captured as humans or social bots, we can associate each user with the veracity levels of the threads which they chose to comment on. It seems natural that there may be a correlation between the status of a user as human or machine and the frequency with which they comment on *mostly false* or *mixture of true and false* posts.

Disqus analysis

An additional subject we investigated during the literature review was any discussion of comments made using the Facebook or Disqus comments plugins. We found no such work. In the case of Facebook, this seems reasonable as the comments made on these third-party sites are functionally identical to those made on the platform itself. On the other hand, we found it extremely surprising to find no work on Disqus, considering it is currently one of the most popular commentary plugins on the internet. While the majority of our dataset is made up of Facebook comments, there is still a sizeable collection of data from Disqus comments. Since there seems to be no scholarly work on this subject, there are quite a few possible avenues for research. An immediate possibility is the analysis of differences between commentary on Disqus and Facebook, as any time our dataset yields Disqus commentary on an article we are sure to also have supplementary Facebook commentary on the same post. Disqus users may represent an entirely separate demographic than the Facebook users in that one needs no social media account whatsoever to sign up for Disqus and begin commenting. Considering the ubiquity of this plugin, the richness of the data their API provides, and the lack of scholarly work on the subject, future studies into Disqus look quite promising.

Methods

Facebook Graph API

The data provided by BuzzFeed came in the form of a CSV file with each row representing a single news item posted to Facebook. The essential feature in each row for our endeavors was the Facebook post ID associated with the post. Python scripts were constructed to loop through all the post IDs given and insert these IDs in the constructed URLs which queried the Facebook Graph API for the features which we desired, including: comments, information about the post itself, the attachments, and the statistics concerning shares. This information was then stored with the appropriate directory hierarchy as discussed previously, with JSON files representing the above mentioned data obtained from the API queries.

Accessing web content

Each Facebook post in the BuzzFeed dataset came labeled with one of the following types: *video*, *link*, or *photo*. Collecting *video* and *photo* posts only required another call to the Facebook Graph API. In order to obtain the body text and other important features of the actual articles themselves (the *link* type), we accessed articles associated with the Facebook posts using Python and the modules BeautifulSoup and urllib2. Given the vastly different styles of web code for the different outlets, at least one tailor-fit utility was required for each outlet.

Before web content could be accessed, another selection process was required for the articles in queue. It is very common for news organizations on Facebook to share articles written by other outlets, so prior to using our tools we had to determine which posts were actually produced by which outlets. For each outlet we looked at several examples of news articles posted and examined their URLs for strings that could be used to identify them. We then set up a hash map relating outlets to these identifier strings, and iterated through each article URL checking for the appropriate identifier string. If the identifier string was present, we considered the article to be first-party. We created a Boolean associated with each article to store this information.

After some analysis of the sources of the articles for the various outlets, it became apparent that this was not enough. Out of *The Other 98%*'s 121 Facebook posts, 51 of them are links to text articles. Out of these links, none of them are articles on *The Other 98%*'s webpage, while a massive 35 (68.63%) of them were *US Uncut*'s (another outlet) articles. At this point, it became clear it would be worthwhile to create a *US Uncut*-specific utility in place of that for *The Other 98%*'s. There are additionally 12 *Occupy Democrats* articles amongst the remainder, which is an outlet we had already established tools for, so it was easily applied to these. The remaining third-party pages with minor representation were simply skipped. These articles along with those that have since been deleted make up the entirety of the articles which were not accessed. Only about 4.05% of the articles have been deleted, while 20.42% are third-party; Table 2 further illustrates these statistics.

Conclusion

We have collected and adjoined to the BuzzFeed dataset a massive amount of additional data. Not only is the size of the data impressive, but our contribution is feature-rich, well organized, and has been made simple to navigate for other users to perform their various analyses. The contribution of over 1.6 million additional pieces of text that are directly related to news items analyzed by BuzzFeed will allow for a truly robust and intriguing dataset. Such a large gold-standard dataset geared towards news veracity assessment has simply not existed before this time, which makes our contribution to the BuzzFeed dataset highly beneficial for this endeavor. Moreover, our timely access to user-level information over the initial integration of *BuzzFace* has allowed us to open a window into the interactions of users on Facebook. Not only have we maintained these extremely important data for our own analyses, but we have anonymized them as Ego-IDs for community access, making this dataset a one of a kind and potent object for the research community. On this note, we also highlight the ephemerality and changing nature of *BuzzFace*. Users are still commenting on the dataset's news articles (albeit more slowly), in addition to deleting some posted content (and now even accounts).

Anyone who wishes to utilize *BuzzFace* as described should act to integrate the dataset now.

Apart from the massive size differential between our dataset and those of the status-quo, it is also important to note that all previous datasets have focused strictly on content alone. The addition of supplementary data such as user comments provides entirely new dimensions. Additionally, the evolution beyond a truth dichotomy to a full four categories of veracity makes for far more subtlety and power in a veracity assignment mechanism. The data also is poised to be helpful for the detection of social bots. The 2016 U.S. Presidential Election is well-known to have been a time where social bot usage was rampant on various social media platforms, particularly Facebook. The dataset we have proposed offers a massive sampling of commentary on articles during this time period where nearly every news item posted by the nine outlets regarded the Election. This is precisely where one may expect social bots designed to alter the outcome of the Election would have been most active. The breadth, along with the timing and placement of our proposed dataset makes it very valuable for the study of both social bots in general and particularly their influence and methods used during the Election. For data development, the next steps are clear: follow the leads to other social media platforms and use their APIs to supplement the dataset. Beyond the two platforms mentioned (Twitter and Reddit), perhaps ties could be found and implemented to spread even further.

References

Center, P. R. 2015. The evolving role of news on twitter and facebook. http://assets.pewresearch.org/wp-content/uploads/sites/13/2015/07/Twitter-and-News-Survey-Report-FINAL2.pdf. Accessed: 2018-01-17.

Chen, Y.; Conroy, N. J.; and Rubin, V. L. 2015. News in an online world: The need for an automatic crap detector. *Proceedings of the Association for Information Science and Technology* 52(1):1–4.

Conroy, N. J.; Rubin, V. L.; and Chen, Y. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1):1–4.

Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Hoi, G. W. S.; and Zubiaga, A. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In Bethard, S.; Carpuat, M.; Apidianaki, M.; Mohammad, S. M.; Cer, D.; and Jurgens, D., eds., *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 69–76. Vancouver, Canada: Association for Computational Linguistics.

Ecker, U. K.; Lewandowsky, S.; Chang, E. P.; and Pillai, R. 2014. The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied* 20(4):323.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Gottfried, J., and Shearer, E. 2016. News use across social media platforms 2016.

Howard, P. N., and Kollanyi, B. 2016. Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. *arXiv*.

Howard, P. N.; Kollanyi, B.; and Woolley, S. 2016. Bots and automation over twitter during the us election. *Computational Propaganda Project: Working Paper Series*.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Rieder, B. 2013. Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th annual ACM web science conference*, 346–355. ACM.

Rubin, V. L.; Chen, Y.; and Conroy, N. J. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology* 52(1):1–4.

Santia, G. C., and Williams, J. R. Buzzface: a news veracity dataset with facebook user commentary and egos. https://dataverse.mpi-sws.org/dataverse/icwsm18. Accessed: 2018-04-10.

Silverman, C.; Strapagiel, L.; Shaban, H.; Hall, E.; and Singer-Vine, J. 2016. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate.

Team, F. N. C. 2018. Exploring how artificial intelligence technologies could be leveraged to combat fake news. http://www.fakenewschallenge.org/. Accessed: 2018-01-08.

Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.

Vlachos, A., and Riedel, S. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Weedon, J.; Nuland, W.; and Stamos, A. 2017. Information operations and facebook. Technical report, Facebook.