SBP-BRiMS 2017 Challenge Project Submission

Title: Automated Linguistic Classification of Fake and Real News

Name	Email	Institution
Grayson Cupit	jgcupit@memphis.edu	*
Leah Windsor, Ph.D.	Leah.Windsor@memphis.edu	*
Danielle Clewley	dnclwley@memphis.edu	*
Michael Boyd	mboyd6@memphis.edu	*
Alistair Windsor, Ph.D.	awindsor@memphis.edu	Department of Mathematical Sciences*

**Institute for Intelligent Systems, The University of Memphis*

Theoretical framework

Why has the issue of fake news gained traction in the United States in recent months, and what functions does it serve?[1] The term 'fake news' refers to multiple phenomena, including the deliberate spread of false information, satire, outdated/revived content, hoaxes, clickbait, propaganda, and disinformation. Recent research on the "50c army" shows that contrived social media posts serve to distract and redirect public narrative during times of crisis or negative publicity, which may be the goals of fake news propagators.[2] The key problem with fake news is determining truth from fiction, a form of deception detection. Readers interact with both headlines and the body of text to discern the veracity of the source. Some fake news detectors rely on human raters, such as the BS Detector, Fake News Alert, and Politifact. FiB and Stop-the-Bullshit[3] are automated tools built for social media, but are not available for the Internet at large. Other scholars have pioneered automated fake news detection by mapping the diffusion pattern of 'likes' and 'shares' automatic hoax detection systems.[4] Facebook has crowdsourced solutions to the problem of identifying fake news, including verified news sites (akin to Twitter's strategy), separating 'shares' from personal information, time delays on 'reshares', Snopes partnership, and headline and content analysis.[5] We focus on automated content analysis to develop a fake news classifier that relies on linguistic features rather than human moderation or social media integration. What linguistic cues do these stories rely on to engage readers and promote interaction and how can we use these cues to automatically differentiate fake from real news?

Data Collection

We use the Kaggle "Getting Real about Fake News" dataset (full texts given) as ground truth for fake news stories. It comprises of 12,999 posts from 244 fake news websites. We remove non-English language entries and entries that lack a headline or full article text, resulting in a usable dataset (n=11,568). For headlines, we use the Buzzsumo service to measure social media engagement (all shares across popular Social Network platforms). We assemble news stories published by reputable news outlets (CNN, Fox News, MSNBC, New York Times, Reuters, and Al Jazeera) to establish a comparison corpus. We use the Buzzsumo service to search for English language articles from any web domains associated with those sites.[6] An HTML scraper is built in Python to capture articles from these URLs. To ensure a balanced distribution of news sources, we randomly select 1,833 articles from each source (n=10,998). We generate four corpora from these datasets: real news headlines, fake news headlines, real news articles, and fake news articles. We generate the real news article corpus by implementing an HTML scraper in Python to retrieve article texts from the URLs returned by Buzzsumo. We discard entries that have no usable article text (e.g. videos) resulting in a usable corpus (n = 6,081).

Method

For real and fake news headline corpora, we analyze document headlines using Linguistic Inquiry and Word Count (LIWC) 2015, resulting in 93 measures per document describing the cognitive, affective, and grammatical processes of the text.[7] We use a truncated singular value decomposition (SVD) to compress each data point to the top 70 singular values, preserving 97.7% of the variance. From this, we perform a t-Stochastic Neighbor Embedding (t-SNE) to assess the separability of the data (Fig. 1).[8] We perform a two-tailed, independent samples t-test between LIWC features of fake and real headlines and find that 68 of 93 measures are significantly different. Fake news headlines use significantly more quotation marks, function words, and conjunctions and less male language (e.g. *boy, his, dad*) on average. We apply an identical methodology to the corpus of full text articles. We perform singular value decomposition, retaining top 70 values, preserving 98.7% of the variance. We use a t-SNE algorithm to embed 70 dimensions into two and we observe a greater degree of separability. This inspires confidence in the ability of traditional classification algorithms to perform well on the dataset. We again perform a two-tailed, independent samples t-test on linguistic features and 80 out of 93 LIWC measures are significantly different. We also find larger effect sizes when considering the article bodies (Table 1). Fake news articles are less reflective and analytic but more certain.

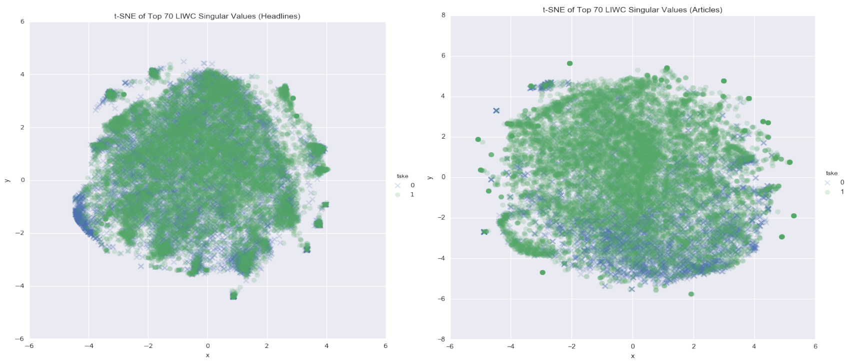

Figure 1. *D t-SNE Plots of LIWC Dimensions (left, headlines; right, articles)*

From the figure, we see clusters of real or fake news headlines, with the majority of them remain unclustered. This demonstrates that in most cases, credibility cannot be usually established from the headline alone.

Table 1.*T-test comparing fake article language to real article language*

p	LIWC measure	Cohen's *d*
0.00	hear	-0.84
0.00	focuspast	-0.75
0.00	focuspresent	0.64
2.96E-200	certain	0.45
2.97E-184	Comma	-0.46
4.16E-168	Colon	0.44
1.98E-159	Analytic	-0.43
5.33E-151	adverb	0.42
3.51E-145	work	-0.41
1.66E-131	Quote	-0.39
1.05E-43	Dic	0.19
7.55E-42	Exclam	0.18
2.39E-35	male	-0.17

Classifier and Techniques

We propose an ensemble method for classifying news articles to predict their credibility as a binary variable (fake vs. real). We also propose a neural network based model to predict the social media success of a news article based on its linguistic features. For both the classification and regression tasks, we propose (and observe the performance of) models utilizing different linguistic features: individual words, LIWC measures, and Latent Dirichlet Allocation (LDA) topic modeling distributions.[9] Finally, we construct and evaluate a mixed model using these feature spaces. To perform classification based on words used, we use the *word2vec* implementation of the vector space embedding process proposed by Mikolov et al., to assign each unique word a unique vector representation within that space.[10] A binary classifier is trained to predict the credibility of news article based on the set of comprising vectors. Standard measures of classifier performance such as AUROC and F1 scores are reported. We then construct several topic models using LDA methods, with the corpus consisting of all fake and real headlines and articles, using a range of possible numbers of topics. The efficacies of the LDA models are evaluated qualitatively based on the topic keywords, and the most effective model is used to assess the topic distribution of individual articles, mapping each article to a *k*-vector, where *k* is the number of topics used to generate the model. Finally, we train a classifier on the results of LIWC analysis before constructing a new model using word vectors, LDA, and LIWC data as features.

Next Steps

Additional next steps include exploring the semantic similarity of fake news headlines and text body.We will also further explore the goals of fake news in the United States from an event data perspective, using vector autoregression to determine Granger causality, using LDA models to examine the evolution of fake versus real news topics over time. We can also assess the 'virality' of fake and real news based on the numbers of 'likes' and 'shares' on social media. Proposed future work involves an experimental approach to evaluate readers' cognition and emotion while reading fake or real news articles, and our research team has the capacity for an experimental design to establish fake versus real classification based on participants' EEG.[11] Distinguishing fake from real news is a substantively important question to engage given the sociopolitical stakes with implications for issues such as low- and high-information voters during increasingly polarized partisan politics not only in the United States, but myriad other mature democracies.

[1] Hunt Allcott and Matthew Gentzkow, “Social Media and Fake News in the 2016 Election” (National Bureau of Economic Research, 2017), http://www.nber.org/papers/w23089; Charlie Beckett, “‘Fake News’: The Best Thing That’s Happened to Journalism,” POLIS: Journalism and Society at the LSE, 2017, http://eprints.lse.ac.uk/76568/; Dan Berkowitz and David Asa Schwartz, “Miley, CNN and The Onion: When Fake News Becomes Realer than Real,” Journalism Practice 10, no. 1 (2016): 1–17; Michael C. Dorf and Sidney Tarrow, “Stings and Scams:‘Fake News,’the First Amendment, and the New Activist Journalism,” 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2906444; Fook Kwang Han, “Can Governments Stop Fake News?,” 2017, https://dr.ntu.edu.sg/handle/10220/42339; Dan Kahan, “The Psychology of Fake News,” in 2017 AAAS Annual Meeting (February 16-20, 2017) (aaas, 2017), https://aaas.confex.com/aaas/2017/webprogrameurekalert/Paper20704.html; Hans Klein et al., “The Rise of Fake News,” March 16, 2017, https://smartech.gatech.edu/handle/1853/56548; Adam Kucharski, “Post-Truth: Study Epidemiology of Fake News,” Nature 540, no. 7634 (2016): 525–525; Regina Marchi, “With Facebook, Blogs, and Fake News, Teens Reject Journalistic ‘Objectivity,’” Journal of Communication Inquiry 36, no. 3 (2012): 246–262; Nicco Mele et al., “Combating Fake News: An Agenda for Research and Action,” 2017, https://shorensteincenter.org/wp-content/uploads/2017/05/Combating-Fake-News-Agenda-for-Research-1.pdf; Gordon Pennycook, Tyrone D. Cannon, and David G. Rand, “Prior Exposure Increases Perceived Accuracy of Fake News,” SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, April 30, 2017), https://papers.ssrn.com/abstract=2958246.

[2] Gary King, Jennifer Pan, and Margaret E. Roberts, “Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation,” Science 345, no. 6199 (2014): 1251722.

3 http://bsdetector.tech/; http://mashable.com/2016/11/16/fake-news-alert-chrome-extension/; www.politifact.com/georgia/subjects/fake-news/

[4] Eugenio Tacchini et al., “Some Like It Hoax: Automated Fake News Detection in Social Networks,” ArXiv:1704.07506 [Cs], April 24, 2017, http://arxiv.org/abs/1704.07506.

[5] Nicky Woolf, “How to Solve Facebook’s Fake News Problem: Experts Pitch Their Ideas,” The Guardian, November 29, 2016, sec. Technology, https://www.theguardian.com/technology/2016/nov/29/facebook-fake-news-problem-experts-pitch-ideas-algorithms.

[6] http://buzzsumo.com/

[7] James W. Pennebaker et al., “The Development and Psychometric Properties of LIWC2015,” *UT Faculty/Researcher Works*, 2015, https://utexas-ir.tdl.org/handle/2152/31333.

[8] Laurens van der Maaten and Geoffrey Hinton, “Visualizing Data Using T-SNE,” *Journal of Machine Learning Research* 9, no. Nov (2008): 2579–2605.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research* 3, no. Jan (2003): 993–1022.

[10] Tomas Mikolov et al., “Efficient Estimation of Word Representations in Vector Space,” *ArXiv Preprint ArXiv:1301.3781*, 2013, https://arxiv.org/abs/1301.3781.

[11] Kai Kunze et al., “Reading Activity Recognition Using an Off-the-Shelf EEG–Detecting Reading Activities and Distinguishing Genres of Documents,” in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference On* (IEEE, 2013), 96–100, http://ieeexplore.ieee.org/abstract/document/6628592/.