11-2018

Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017)

Victoria Rubin
University of Western Ontario, Faculty of Information and Media Studies, vrubin@uwo.ca

Toluwase Victor Asubiaro
Western University, tasubiar@uwo.ca

Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017)

Toluwase Victor Asubiaro
University of Western Ontario, Canada. tasubiar@uwo.ca

Victoria L. Rubin
University of Western Ontario, Canada. vrubin@uwo.ca

ABSTRACT

With the problem of 'fake news' in the digital media, there are efforts at creation of awareness, automation of 'fake news' detection and news literacy. This research is descriptive as it pulls evidence from the content of online fabricated news for the features that distinguish fabrications from the legitimate political news around the time of the U.S. Presidential Elections (276 articles in total, from November 2016 - June 2017). Certain stylistic and psycho-linguistic features of fabrications may be apparent to the news readers: fewer words and paragraphs but longer paragraphs, more slangs, swear words and affective words in the stories. Such features could be used for educational information literacy campaigns for spotting so-called 'fake news'. Other informative features may require specialized analytical tools (or further training) to notice the presence of more words, punctuation marks, demonstratives and emotiveness in fabrications but fewer verifiable facts (or named entities) in their headlines.

KEYWORDS

News fabrication, Fake news, Disinformation, Deception, Content analysis, Information literacy.

INTRODUCTION

Information literacy (IL) includes a set of skills that require individuals to "have the ability to locate, evaluate, and use effectively the needed information" (American Library Association, 1989). In recent times, the ability to evaluate information from news sources is a new challenge to the news readers in the digital environment because of the proliferation of fabricated news, often referred to as 'fake news'. In response to the need for IL programs to incorporate skills for identifying misinformation and disinformation online, there are new IL programs which are customized towards 'fake news' identification (International Federation of Library Associations, 2018; Wyman, 2017).

In the Tandoc, Lim, & Ling (2018) 'fake news' typology, news fabrication (or falsification) is defined as having low facticity with high author's intention to deceive. In other words, the content of news fabrication is not factual and it is presented with the intention to create a false impression or conclusion in the reader's mind. Studies have shown that humans are not effective at the lie-truth discrimination task; typical accuracy rates are in the 55–58% range, with a 54% mean accuracy of over 1,000 participants in over 100 experiments (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997). On the other hand, some algorithmic applications, called automated deception detection, can be more accurate than humans in certain contexts (Rubin & Conroy, 2012). Automatic deception detection for the news context is hinged on contrasting differences in the linguistic, stylistic and psycholinguistic features between the legitimate and fabricated news articles. The differences in some of these features are not easily observed or recognizable by humans, for instance, pronoun or punctuation marks frequencies, but can be tallied and monitored algorithmically. As we are doing research in preparation for automation of identification of fabricated news, we perform comparative statistical analysis of linguistic features of fabricated versus legitimate news. Despite substantial difficulties expected for news readers in separating legitimate news from fabricated ones, news readers could and should pay attention to certain statistically significant differences between the two types of texts that we identified in our dataset.

The aim of this research is to highlight some human-noticeable features which can help human differentiate fabricated news from legitimate ones. The idea is that some of these features are informative, and they can be taught and incorporated into the IL programs for spotting 'fake news' with a naked eye. For an obvious example, consider the use of customary more formal language in news articles. Overt incessant non-standard language use, such as slang and swearing, should be raise a red flag.

METHODOLOGY

We collected 276 digital articles in the domain of the U.S. politics in order to find differences between fabricated (i.e., 'fake') and legitimate news, emanating from the United States between November 2016 and June 2017. The volume of fabricated news and the associated problems became unprecedented around the time of the 2016 U.S. Presidential Elections. Three steps were taken during our data collection. First, the PolitiFact.com was examined for the 50 most current websites identified at the time of data collection as 'fake news' sources with 'pants-on-fire' ratings. Secondly, we collected 5 most up- to-date fabricated (or falsified) news articles from the sources, identified via PolitiFact. The resulting total was 185 'fake news' from 37 websites relating directly to politics. Out of those, 138 falsifications were used for direct matching with legitimate news; we ended up with 138 pairs of news articles (138 fabricated and 138 legitimate, 276 articles in total). We manually matched the falsifications

to their potential inspirations. We assumed high probability of falsifications emanating from some legitimate news and limited the dataset to U.S. politics only. Each article and headline was carefully read (by the first author) and key words were extracted manually to reflect the theme or the subject. Keyword search strategy was used with the Google search engine and the Reuters database and verified by a close reading for the best match between the fakes and their matching legitimate news.

We content-analyzed the matched datasets of 276 fabricated and legitimate news headlines and texts with natural language processing (NLP) techniques, using *pattern.en* (De Smedt, and Daelemans, 2012) and the NLTK packages (Loper & Bird, 2002) of the Python language libraries. The following features were collected: word count per news story, paragraph and headline; number of affect words, number of informalities (swear words and slangs), and verifiable facts (referred to in NLP as named entities which can be proper names of individuals, things, places, times and dates). Other features that were extracted are affect (positive and negative wording), emotiveness (counts of adverbs + adjectives divided by counts of noun + verbs) and frequencies of demonstratives ('this', 'that', 'these') and pronouns ('he', 'she', 'mine', 'hers', *etc*.) Paired sample *t*-test was conducted on pairs of legitimate and fabricated news with significant level set at 0.05.

RESULTS AND DISCUSSION

The result of the paired sample *t*-test is presented in Table 1 where each pair (from 1 to 10) refers to legitimate versus fabricated news pair. Let us consider differences in the surface stylistic (lexical) features first. Legitimate news stories contain an average of 481.87 more words ($F(1, 137)=5.88$, $p\approx0.01$) than fabricated news. Similarly, legitimate news contain an average of 14 more paragraphs ($F(1, 137)=6.90$, $p\approx0.01$) than fabricated news. On the other hand, legitimate news contain shorter paragraphs ($F(1, 137)=-3.00$, $p=0.04$), an average of 8.06 number of words per paragraphs less than fabricated news that contain longer paragraphs. Figure 1 (see Appendix) shows a sample each from legitimate and fabricated news sources where the legitimate news has shorter paragraphs and bigger fonts, while the fabrications are presented with smaller fonts but longer paragraphs. On the other hand, fabricated news headlines contain more words ($F(1, 137)=3.00$, $p=0.03$) and more punctuation marks ($F(1, 137)=-4.06$, $p\approx0.01$) than the sample legitimate news headlines. It may be difficult for humans to keep track of punctuation marks in the body of the news but spotting unnecessary punctuation marks in the headline is trivial.

No	Features Compared	Mean	Std. Dev	Std. Error Mean	95% Confidence Interval of the Diff		*t*	Sig. (2-tailed)
					Lower	Upper		
1	No of words/story	481.87	962.23	81.91	319.90	643.84	5.88	.000
2	No of paragraphs/story	14.09	24.00	2.04	10.05	18.13	6.90	.000
3	No of words per paragraph/story	-8.06	32.03	2.73	-13.46	-2.67	-3.00	.004
4	Affect/story	-.20	.17	.02	-.23	-.17	-13.66	.000
5	Informality/story	-.02	.06	.005	-.03	-.01	-4.26	.000
6	No of words/headline	-1.74	6.84	.58	-2.89	-.59	-3.00	.003
7	Verifiable facts/headline	1.10	.97	.08	.94	1.27	13.36	.000
8	No of punctuations/headline	-.25	.71	.06	-.37	-.13	-4.06	.000
9	Demonstratives/headline	-.25	1.02	.09	-.42	-.07	-2.82	.005
10	Emotiveness/headline	-.06	.04	.003	-.07	-.06	-20.17	.000
The *df*=137								

Table 1. Paired Sample *t*-test result of Legitimate vs. Fabricated News

Differences in psycho-linguistic features also show that fabricated news articles contain more positive and negative affect ($F(1, 137)=-13.66$, $p\approx0.01$) and their headlines contain more emotiveness ($F(1, 137)=-20.17$, $p\approx0.01$) showing attempts at heavy emotional appeals to the readers in the bodies of such articles and their headlines. We found more informal words in fabricated news stories ($F(1, 137)=-4.26$, $p\approx0.01$), as expected. Also, fabricated news headlines contain more demonstratives (pronouns and unspecific) in fabricated news headlines ($F(1, 137)=-13.66$, $p\approx0.01$) and on the other hands, less verifiable facts (specific names) ($F(1, 137)=13.36$, $p\approx0.01$) as often seen in clickbait. In essence, put in plain language, fabricated news headlines contain 'more of *he, she, they, etc*.' (i.e., pronouns), while legitimate news headlines contain more specific names. Figure 2 shows fabricated vs. legitimate news headlines (see Appendix) with more words, demonstratives, pronouns, and punctuations marks but with fewer verifiable facts (or named entities).

CONCLUSIONS

Based on the results of the paired *t*-test between fabricated and paired legitimate news in the U.S. politics from 2016-2017, we identified several features that can be incorporated in the news literacy awareness campaigns targeting broader awareness on how to spot 'fakes' in the digital news environments. Fabricated political news stories by comparison to their likely legitimate counterparts, tend to have fewer words, fewer but lengthier paragraphs; they also contain more slang, swear, and affective

words. The fabricated news headlines contain more words, punctuation marks, demonstratives, emotiveness and fewer verifiable facts.

ACKNOWLEDGMENTS

This research has been funded by the Government of Canada Social Sciences and Humanities Research Council (SSHRC) Insight Grant (#435-2015-0065) awarded to Dr. Rubin for the project entitled *Digital Deception Detection: Identifying Deliberate Misinformation in Online News.*

REFERENCES

American Library Association. (1989). *Presidential Committee on Information Literacy.* (Final Report). Retrieved from www.ala.org/Template.cfm?Section=Home&template=/ContentManagement/ContentDisplay.cfm&ContentID=33553#f1

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J. & Muhlenbruck, L. 1997. The Accuracy-Confidence Correlation in the Detection of Deception. Personality and Social Psychology Review, 1, 346-357.

De Smedt, T. & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13: 2031–2035.

International Federation of Library Associations. (2018). How to spot fake news. Retrieved March 29, 2018, from https://www.ifla.org/publications/node/11174

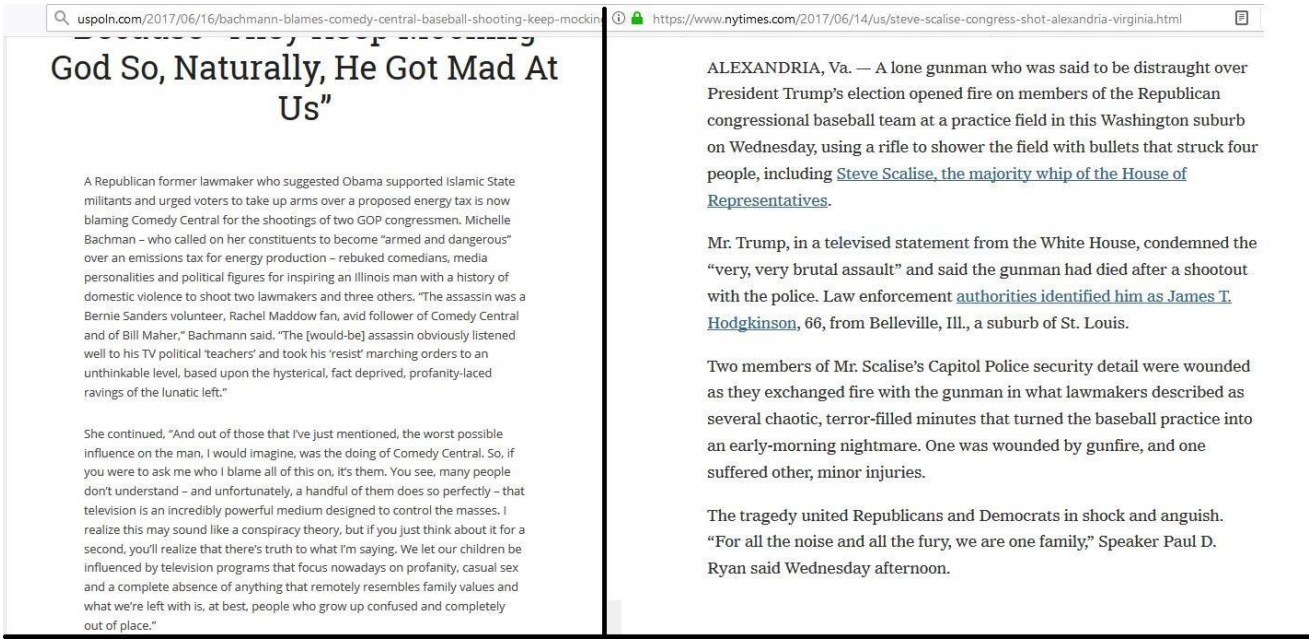
Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. In *ETMTNLP '02 Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* (Vol. 1, pp. 63– 70). Philadelphia, Pennsylvania: Association for Computational Linguistics.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Rubin, V. L., & Conroy, N. (2012). Discerning truth from deception: Human judgments and automation efforts. *First Monday*, *17*(3). https://doi.org/10.5210/fm.v17i3.3933

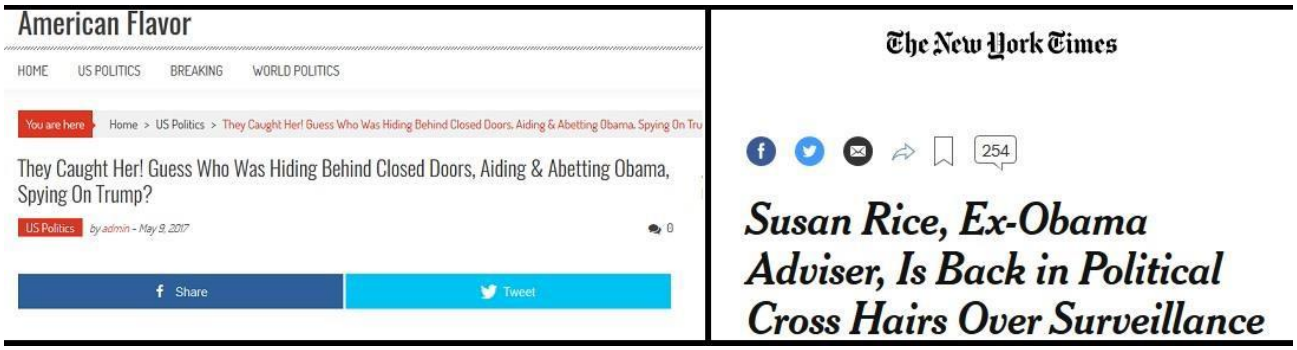
Wyman, J. (2017). How to Identify Fake News in 10 Steps. Retrieved March 29, 2018, from http://blogs.proquest.com/general/how- to-identify-fake-news-in-10-steps/

APPENDIX



a. An example of fabricated news story. **b.** An example of legitimate news story.

Figure 1. Fabricated versus Legitimate news story



a. An example of fabricated news headline. **b.** An example of legitimate news headline.

Figure 2. Fabricated versus Legitimate news headline.