**SATIRICAL AND LEGITIMATE NEWS DATABASE (S-n-L NEWS DB2015-2016)**

This **README file** is a description of a database of satirical and legitimate news articles (S-n-L NEWS DB 2015-2016) collected from September 2015-January 2016 and used in Rubin et al NAACL-CADD2016 research article.

**License:** The S-n-L NEWS DB 2015-2016 is an Open Educational Resources (OER) and is now released under an open license for educational and research purposes.

**Password:** The DB site is password-protected but we will readily share the password with you upon your request. (Not much luck for automated site crawlers.)

**Citation:** Rubin, V.R., Conroy, N. J., Chen, Y. & Cornwell, S. (2016) Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In the Proceedings of the Workshop on Computational Approaches to Deception Detection at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-CADD2016), San Diego, California, June 17, 2016.

The S-n-L NEWS DB 2015-2016 was collected by the Language and Information Technology Research Lab (LiT.RL) directed by Dr. Victoria Rubin, Western University, London, Ontario, Canada, <http://victoriarubin.fims.uwo.ca/research/>

This collection was part of the 2015-2018 News Verification Project funded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

**Collection Method:**

The 360 news articles were collected and analyzed to create a wide-ranging and diverse data sample, representative of the scope of US and Canadian national newspapers. The dataset was collected in 2 sets.

**SET 1:** The first 240 articles (published in 2015) were aggregated into a 2x2x12 design (US and Canadian; satirical and legitimate online news; varying across 4 domains (civics, science, business, and “soft” news) with 3 distinct topics within each domain (see labels in Column E). The 240 news pieces were carefully selected with an equal representation (5 articles per subtopic listed in Column E).

**SET 2:** An additional set of 120 articles was collected from online publications in 2016, to expand the inventory of sources and topics and to serve as a reliability test for the manual findings within the first set, the second set was still evenly distributed between satirical and legitimate news.

**Dataset Details:**

**SET 1:** The 2015 data for this project were articles collected from 2 satirical news sites (The Onion and The Beaverton) and 2 legitimate news sources (The Toronto Star and The New York Times). The publications were also paired by country of origin, with The Onion and The New York Times originating in the United States, and The Beaverton and The Toronto Star originating in Canada. For each of the 12 topics, 5 Canadian (from The Beaverton) and 5 American (from the Onion) satirical articles were collected. Each satirical piece was then matched to a legitimate news article that was published in the same country, and as closely matched in subject matter as possible. For example, in the Environment topic, the Beaverton article “Invasive homo sapiens species meet at forestry conference to discuss pine beetles” was paired with a Toronto Star article on invasive species: “'Dangerous and invasive' Khapra beetle intercepted at Pearson”. Similarly, for the Elections/Debates topic, the Onion’s article “Campaign Staffers Making Progress Conditioning Hillary Clinton To Replicate Emotions” was paired with the New York Times’ “Hillary Clinton to Show More Humor and Heart, Aides Say”. See illustrations in Figure 2 for screenshots of the news matched on topics.

**SET 2:** The second dataset was more varied in topics and was drawn from 6 legitimate and 6 satirical North American online news sources which were active and frequently updated in 2015.

**Database Description:**

Column A: Satirical news pieces are marked with 1, legitimate news pieces are marked with 0

Column B: The article’s headline

Column C: The full text of the article, no headline

Column D: Domain\*. (Politics/Civics, Soft News, Health/Science, Business/Technology).

\*Domain only applies to articles from The New York Times, The Toronto Star, The Onion, and The Beaverton

Column E: Subtopic. Subtopic only applies to articles from The New York Times, The Toronto Star, The Onion, and The Beaverton

Politics/Civics: Immigration, Election, Gun control

Immigration: Immigration issues, refugees, etc.

Election: Canadian Election 2015, US Election 2016 Debate and politician coverage

Gun Control: Mass shootings, terrorism

"Soft" News: Sports, Entertainment, Local News

Sports: PanAm Games, sporting events, sport celebrities

Entertainment: celebrity news and scandals

Local News: local events, personal interest stories

Health/Science: Hard Science, Health, Environment

"Hard" Science: astronomy, physics, chemistry, biology, palaeontology, etc.

Health: Health, medicine, public health, epidemiology

Environment: Climate change, endangered species, environmental biology

Business/Technology: Tech Improvements, Banking/Finance, Business Announcements

Tech improvements: Technology sector news, tech trends, the future of technology, futurology

Banking/Finance: Economy, Banking, stock markets

Business Announcements: business press releases, non-technology sector trends

Column F: Publication Title

**Satirical: Main (set 1): US:** The Onion **Canada:** The Beaverton, **Other (set 2) : Canada:** CBC’s Punchline, The Codfish, The Lapine, The Syrup Trap, Sage News **US:** The Daily Currant, The Spoof, The National Report, The People’s Cube, World News Daily Report, Urban Anomie

**Legitimate: Main (set 1): US:** The New York Times **Canada:** The Toronto Star **Other (set 2): Canada:** The Globe and Mail, The Vancouver Sun, The Calgary Herald, The National Post, The Edmonton Journal, The Hamilton Spectator **US:** USA Today, The Wall Street Journal, The Los Angeles Times, The New York Post, Newsday, The Denver Post

Column G: Date Article published (MM-DD-YYYY)

Column H: Article URL

Column I: Date the article was initially accessed (MM-DD-YYYY)

**Questions?** Contact: Victoria Rubin: **vrubin@uwo.ca**

*Last update: April 4th, 2016*