Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News

[Please cite the WWW'18 version of this paper]

Benjamin D. Horne
Rensselaer Polytechnic Institute
Troy, New York, USA
horneb@rpi.edu

William Dron
Raytheon BBN Technologies
Cambridge, Massachusetts, USA
will.dron@raytheon.com

Sara Khedr
Rensselaer Polytechnic Institute
Troy, New York, USA
khedrs@rpi.edu

Sibel Adalı
Rensselaer Polytechnic Institute
Troy, New York, USA
adalis@rpi.edu

ABSTRACT

Today, journalist, information analyst, and everyday news consumers are tasked with discerning and fact-checking the news. This task has became complex due to the ever-growing number of news sources and the mixed tactics of maliciously false sources. To mitigate these problems, we introduce the The News Landscape (NELA) Toolkit: an open source toolkit for the systematic exploration of the news landscape. NELA allows users to explore the credibility of news articles using well-studied content-based markers of reliability and bias, as well as, filter and sort through article predictions based on the users own needs. In addition, NELA allows users to visualize the media landscape at different time slices using a variety of features computed at the source level. NELA is built with a modular, pipeline design, to allow researchers to add new tools to the toolkit with ease. Our demo is an early transition of automated news credibility research to assist human fact-checking efforts and increase the understanding of the news ecosystem as a whole. To use this tool, go to http://nelatoolkit.science

KEYWORDS

content analysis, news credibility, fact-checking assistance

1 INTRODUCTION

Understanding and analyzing the news landscape has became a priority for researchers across many disciplines. The production and consumption of news in today's media landscape favors clicks and attention, as opposed to in-depth analysis. This drive for attention has lead to the emergence of a large number of media sources with ever increasing visibility. These sources operate under different incentives: from benign to opportunistic and malicious. Those sources which are partisan or malicious in intent employ a wide-range of tactics to make their message heard. They employ tactics such as reporting incorrect information, using emotionally charged language, manipulative titles, and mixing true news with fake news. Fake news stories and hyper-partisan news coverage are thought to have influenced various key elections worldwide. This, coupled with the well-known susceptibility of individuals to false and misleading information [5], has lead to the increasing need for tools that assist researchers, journalists, and every day individuals in the analysis of news. Supporting this notion, in a 2017 agenda for fake news research, Lazer et al. argue that we "need to translate existing research into a form that is digestible by journalist and public-facing organizations [4]." However, given the complexity, the problem requires multi-faceted solutions and a better understanding of the wide-range of news sources. In addition, tools should be able to quickly evaluate sources to decide where to dedicate fact-checking efforts (before an article's spread).

To address these problems, we introduce the The News Landscape (NELA) Toolkit: an open source toolkit for the systematic exploration of the news landscape, through a unique combination of (a) real data from news sources and social media, (b) state-of-the-art tools that predict different factors of credibility, and (c) visualization tools to compare a large number of media sources across different axes. Specifically, NELA is made up of multiple independent modules, in which users can scrape news articles for article-level predictions or explore source-level characteristics using the built-in NELA data set. In this demonstration, we discuss the first release of the toolkit, and briefly discuss its utility using an initial 7 months of news data from 92 sources across the reliability and bias spectrum.

2 DESCRIPTION OF THE DEMO

To use the NELA Toolkit, visit the NELA Toolkit website (nelatoolkit.science). The homepage provides two choices "Check a News Article" or "Compare News Sources."

Under "Check a News Article" users can provide a url to a news article or manually enter news article text. The tool then performs several predictions on the article: reliability, political impartiality, title objectivity, text objectivity, and several online community interest predictions. Each of these predictions is displayed as a probability and each article with associated predictions are entered into a table. As more article entries are provided, this table can be sorted and filtered by different predictions using the table filters menu at the top of the page. Further, more details about the article and analysis of the article can be found by clicking on the entry in the table. The ultimate goal of this page is to allow journalist and information analyst to quickly filter articles down to ones that need to be fact-checked or are of interest.

Under "Compare News Sources" users can explore and compare a variety of news sources using content-based features. Specifically,

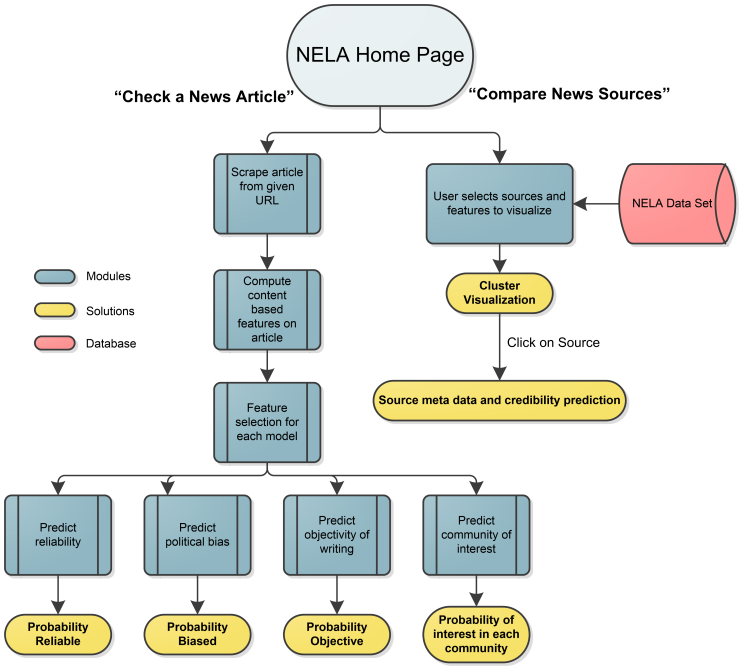


Figure 1: NELA Toolkit architecture

users can select multiple features, sources, and a time range to visualize on a 2-dimensional scatter plot. For example, a user can select "reading complexity" for the x-axis and "negative sentiment" for the y-axis using the chart setting menu on the left side of the page. They can then select any number of sources from our data set and a data range over which to explore. The tool will then generate a scatter plot of the selected sources for comparison. If a user wants more details about a source, they can double-click the source bubble in the scatter plot. This detailed page will show source metadata, credibility predictions, and Facebook engagement over time. These details can also be found on the "View All Sources" page.

The overall architecture of the toolkit can be found in Figure 1. Due to lack of space and the many parts of the toolkit, we do not provide screenshots. We encourage readers to visit the NELA Toolkit website (nelatoolkit.science), watch our demo walk-through (nelatoolkit.science/help), or check out our code-base (goo.gl/cSpWmp

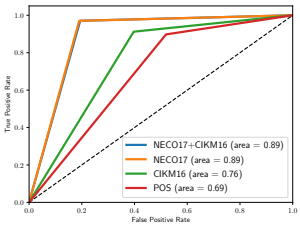
3 DATA

Every module in the NELA toolkit is based on real news data. To create a general news data set, we first gather a wide variety of sources using multiple lexicons (opensources.co, Wikipedia) and studies [2]. These news sources include: mainstream sources, satire sources, maliciously false sources, political blogs, and some relatively unknown sources. Each news source's website or RSS feed is scraped twice a day, everyday, between April 2017 and October 2017, totalling in 92 sources and 136K articles. To control for topic, we only collect news from politics pages and feeds. The complete list of sources currently in the data set can be found on the NELA toolkit website.

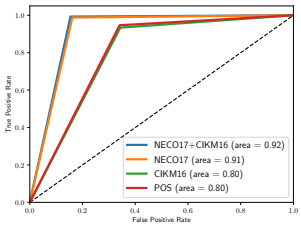
From this general news data set, two subsets are created to build a reliability labeled news data set and a bias labeled news data set. Specifically, we use OpenSources (www.opensources.co/), an expert-curated news source lexicon, to create 4 groups of sources: reliable news, unreliable news, biased news, and unbiased news (Table 2). Opensources has 12 different tags: fake, satire, extreme bias, conspiracy, rumor, state, junk science, hate speech, clickbait, unreliable, political, and reliable. We use the fake and conspiracy tags to create our unreliable group and the bias and political tags to create our biased group. The articles from each labeled source are used in training and testing the two machine learning models, discussed in Sections 4.1 and 4.2.

It is important to note this ground truth is a *previous behavior*-based ground truth rather than a *correctness*-based ground truth. In other words, if a news source has been found to publish many fake articles in the past, they are an unreliable source, or if a news source has been found to be hyper-partisan many times in the past, they are a biased source. We choose this method for two primary reasons: (1) reliability and bias can be labeled quickly over time, allowing for our tool to be retrained as the news changes. Currently, fact-checking (or biased-checking) articles is a very slow and selective process. Hence, fact-checked data for algorithm training can be very small and time specific, making trained classifiers difficult to maintain over time. (2) We can reasonably classify fake articles using this method. Explicitly, on a small fact-checked, correctness labeled test set (of 100 articles), the reliability labeled classifier performs well in detecting fake news as unreliable and real news as reliable (with 90% accuracy). However, our predictions are built to predict the "type of source" a news article is coming from, not the specific nature of the claims in an article. This notion is further discussed in Section 4.

This data will continue to be collected for use in the toolkit and its later release.



Reliable vs. Unreliable



Unbiased vs. Hyper-partisan

Table 1: ROC curves for each feature set using a Random Forest machine learning model, where NECO17 is from [2], CIKM16 is from [8], and POS is a standard Part-Of-Speech feature set.

4 MODULES IN THE NELA TOOLKIT

In this section, we will briefly discuss the basic research behind each module in the NELA Toolkit.

4.1 Reliability prediction

The first module predicts the reliability of a user-selected news article. Given a url, the tool scrapes the title and body content from the web page. After the news article is scraped, it is passed through a feature computation pipeline, which computes a large set of content-based features. These features primarily come from [2, 8], but are also influence by other studies on persuasion [7]. Due to space restrictions, descriptions of these features can be found on the NELA Toolkit website. After features are computed, they are passed through a feature selection module, which selects the best features for the reliability prediction based on a previously computed variance analysis. Once feature selection is done, the single feature vector, representing the user-selected article, is passed to our machine learning model. The reliability model is a Random Forest classifier trained on news sources labeled by previous behavior, discussed in section 3. To make the ground truth stronger, we also require news sources in the unreliable category to have published more than 1 completely false article according to online fact checkers (eg. snopes.com, politifact.com, etc.). In the current implementation, we trained the classifier on 4504 articles and tested it on 1130 articles, achieving 0.89 ROC AUC (refer to Table 1).

The final output of the classifier is a probability of being reliable rather than a strict binary classification. To do this, we use the mean predicted class probabilities from the trees in the forest. This probability is then colored based on the strength of the prediction (where green is strongly reliable, red is strongly not reliable, and yellow is an edge case). This design choice allows for some notion of certainty or uncertainty in the algorithms predictions. News is inherently not a two-class problem, rather a spectrum between the two-classes; hence, it is important to show the user when a data point is near the edge of the decision boundary. Each result is entered into a sort-able and filterable table to allow for batch article analysis. For example, if an analyst is given a large number of news articles to assess, they can use the NELA Toolkit to quickly filter down to the most interesting articles.

4.2 Bias and subjectivity prediction

The next module is made up of two independent classifiers: (1) a Random Forest classifier trained on content-based features to predict hyper-partisan articles, (2) a Naive Bayes classifier trained on objective and subjective labeled sentences. Just as in the reliability module (Section 4.1), a user provides a url, and the title and body content is scraped from the web page. The content is then passed through both feature computation and model-specific feature selection pipelines.

The first classifier in this module is very similar to our reliability module, only differing in the data and features selected. The features are based on several studies on news and political bias in text [2, 9] and the labeled data is discussed in Section 3. The sources are balanced between politically right and politically left hyper-partisan sources. In the current implementation, we trained the classifier on 6158 articles and tested it on 1539 articles, achieving 0.92 ROC AUC (refer to Table 1). The final output from this classifier is a probability of an article being classified as impartial.

The second classifier in this module is more generic than the previous, focusing on sentence level objectivity. Specifically, the classifier will provide a probability of being objective for both the title and body of the news article independently. The separation of title and body allows for a finer-grain analysis of title dynamics. This classifier is built using a Naive Bayes model that is trained on 10K sentences from Pang and Lee 2004 [6], and it achieves a 92% 5-fold cross-validation accuracy. The final outputs of this classifier are the probability of being objective for both the title and body text.

The results from both classifiers are also added to the sort-able and filterable table for quick batch analysis.

Reliable/Unbiased sources	**Unreliable sources**	**Hyper-partisan sources**
Associated Press	Infowars	Brietbart
PBS	Liberty News	Young Cons
NPR	Natural News	RedState
CBS	Alt Media Syndicate	The Blaze
USA Today	DC Clothesline	CNS
BBC	Newslo	Bipartisan Report
New York Times	Ending the Fed	Occupy Democrats
The Guardian	Daily Buzz Live	Daily Kos
	Intellihub	Shareblue
	Freedom Daily	Politicus USA

Table 2: Sources used in each category

4.3 Community interest prediction

Our next module is built to predict which online groups are interested in an article using news communities on `reddit.com`. To build this module, we first collect recent posts from 4 news communities (`r/new_right`, `r/esist`, and `r/conspiracy`). Once these posts are collected, we extract the top 25% of posts by their ranking score (roughly upvotes minus downvotes). These posts can be considered the most popular or most widely accepted by the community during the time slice collected. The news article in each post is scraped and content-based features are computed [3]. We compare `r/news` (a general interest community) to the other three subreddits (specific interest communities). Specifically, using these features, we train 3 binary classifiers to predict articles as "`r/news` interest" or "(`r/new_right`, `r/esist`, `r/conspiracy`) interest." Each classification is shown as a probability, similar to the other modules in the toolkit. In the current implementation, we trained each classifier on 2000 articles and tested each on 500 articles, achieving 0.77 ROC AUC on average.

These community interest models are in a very early stage of development. Currently these models are based solely on news content features, but could be significantly improved with topic, source, or community-specific features. In addition, more in-depth feature analysis can provide insights into community differences and similarities. For example, it may be that highly emotional or subjective articles are popular in both `r/new_right` and `r/esist`, but the articles differ in slant (due to selection bias, framing bias, etc.). Automatic methods to capture these various types of bias in a general news setting could significantly improve our accuracy. We leave these improvements to future work.

4.4 Feature-based source visualizations

Our last module analyzes the news at a source-level granularity, rather than an article-level granularity. Using our data set (refer to Section 3), we computed 260 content-based features [1, 2, 8, 9, 11] on each article. Users can pick a set of news sources, a time frame, and 2 to 4 features to visualize on a 2-dimensional plane. This visualization provides a quick and easy comparison of individual sources or clusters of sources.

Further, we provide meta data for each source, which can be accessed by clicking on a source bubble in the visualization. The meta data includes:

(1) Percentage of articles that were predicted as reliable using our reliability model
(2) Percentage of articles that were predicted as impartial using our bias model
(3) Top phrases for each month using Autophrase [10]
(4) The year the source was founded and the country of origin, if known
(5) Facebook shares, reactions, and comments over time

As data is collected, this module will be updated to reflect the current predictions and articles from each source, allowing for users to explore changes in sources over time.

5 ACKNOWLEDGMENTS

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

6 AUTHORS

Benjamin Horne is a PhD student in Computer Science at Rensselaer Polytechnic Institute. His research focuses on online information quality and credibility, along with the human decisions in assessing online information. This work utilizes techniques in machine learning, natural language processing, and social network analysis to both characterize and detect the veracity information.

William Dron is a Senior Scientist at Raytheon BBN Technologies, where he has worked since 2004. William has a background in communications networks and computer science, particularly for military use. Over the past 8 years, he has focused on cross network genre experimentation and has incorporated social and information sciences in his work. He is particularly interested in utilizing software engineering principles and visualizations to create cohesive and scalable software solutions.

Sara Khedr is a Master's student in the Computer Science department at Rensselaer Polytechnic Institute (RPI). She received her Bachelor's degree in Computer and System Engineering from RPI. During her education, Sara has focused on attaining general software development experience and has previously interned for Workday and Salesforce, two SaaS companies.

Dr. Sibel Adalı is a Professor and Associate Head of Computer Science at Rensselaer Polytechnic Institute. Her research concentrates on cross-cutting problems related to trust, information processing, and social networks. As part of her work, she has worked as the ARL-lead Network Science Collaborative Technology Alliance (NS-CTA) wide Trust Coordinator, Social and Cognitive Networks Academic Research Center (SCNARC) Associate Director.

REFERENCES

[1] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *ASONAM 2016*. IEEE, 9–16.
[2] Benjamin D Horne and Sibel Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. (2017).
[3] Benjamin D Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the social signals that drive online discussions: A case study of Reddit communities. (2017).
[4] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy* 2 (2017).
[5] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131.
[6] Bo Pang and Lillian Lee. [n. d.]. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL 2004*.
[7] R. E. Petty and J. T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *In Communication and Persuasion*. New York: Springer, 1–24.
[8] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. [n. d.]. Credibility assessment of textual claims on the web. In *CIKM 2016*. ACM.
[9] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language.. In *ACL (1)*. 1650–1659.
[10] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2017. Automated Phrase Mining from Massive Text Corpora. *arXiv preprint arXiv:1702.04457* (2017).
[11] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.