NELA-GT-2022: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles

Maurício Gruppi[*], Benjamin D. Horne[†] and Sibel Adalı[*]
Rensselaer Polytechnic Institute[*], The University of Tennessee Knoxville[†]
gouvem@rpi.edu, bhorne6@utk.edu, adalis@rpi.edu

Abstract

In this paper, we present the fifth installment of the NELA-GT datasets, `NELA-GT-2022`. The dataset contains 1,778,361 articles from 361 outlets between January 1st, 2022 and December 31st, 2022. Just as in past releases of the dataset, `NELA-GT-2022` includes outlet-level veracity labels from Media Bias/Fact Check and tweets embedded in collected news articles. The NELA-GT-2022 dataset can be found at: `https://doi.org/10.7910/DVN/AMCV2H`

1 Introduction

Many disciplines utilize news media data in their research, ranging from the study of mass media in journalism to the building of automated tools in computer science. Across these areas, researchers need historical article data that is consistent across time and covers many different types of news outlets. In specific studies of 'fake news' detection, large news datasets with veracity labels are needed.

To these many ends, researchers have focused on building high-quality news datasets. There are several platforms dedicated to collecting news data, such as Media Cloud, an open source platform used for collecting and analyzing global news coverage (Roberts et al. 2021), and LexisNexis, a commercial news database often used in academic studies (Deacon 2007). There are also many one-time news data collections. For example, the FA-KES dataset (Salem et al. 2019), the Golbeck et al. dataset (Golbeck et al. 2018), and the Election-2016 dataset (Bode et al. 2020; Bozarth and Budak 2020). Other one-time datasets focus on social media posts rather than news articles, such as the FakeNewsNet dataset (Shu et al. 2018).

While all of these data sources have been useful for a variety of research studies, there continues to be a need for updated news data. Platforms like Media Cloud do an excellent job at capturing high-quality, current news coverage around the world, but do not capture low-veracity news outlets. Datasets like the Golbeck et al. dataset and the FA-KES datset capture low-veracity news, but quickly become outdated. The yearly-released NELA-GT datasets continue to fill both these gaps: updated news coverage across both low and high veracity outlets.

In this short paper, we describe the fifth release of the NELA-GT datasets, `NELA-GT-2022`. In `NELA-GT-2022` we have collected **1,778,361 articles** from **361 outlets** between **January 1st, 2022 and December 31st, 2022**. Included with these news articles are outlet-level veracity labels from Media Bias Fact Check, with **337 of 361 outlets labeled**, and data on **346,283 distinct tweets** embedded into collected news articles.

In this paper, we describe what is new in the 2022 version of the dataset, the collection process, the publicly-available data formats, and potential use cases.

2 What's New in NELA-GT-2022?

Again, our focus this year was to stabilize our collection infrastructure to ensure complete coverage of articles published across the full year, rather than add new features to the dataset. Hence, as shown in Figure 1, we estimate that our collection has little to no missing article data in 2022.

3 Data Collection

3.1 News data and metadata

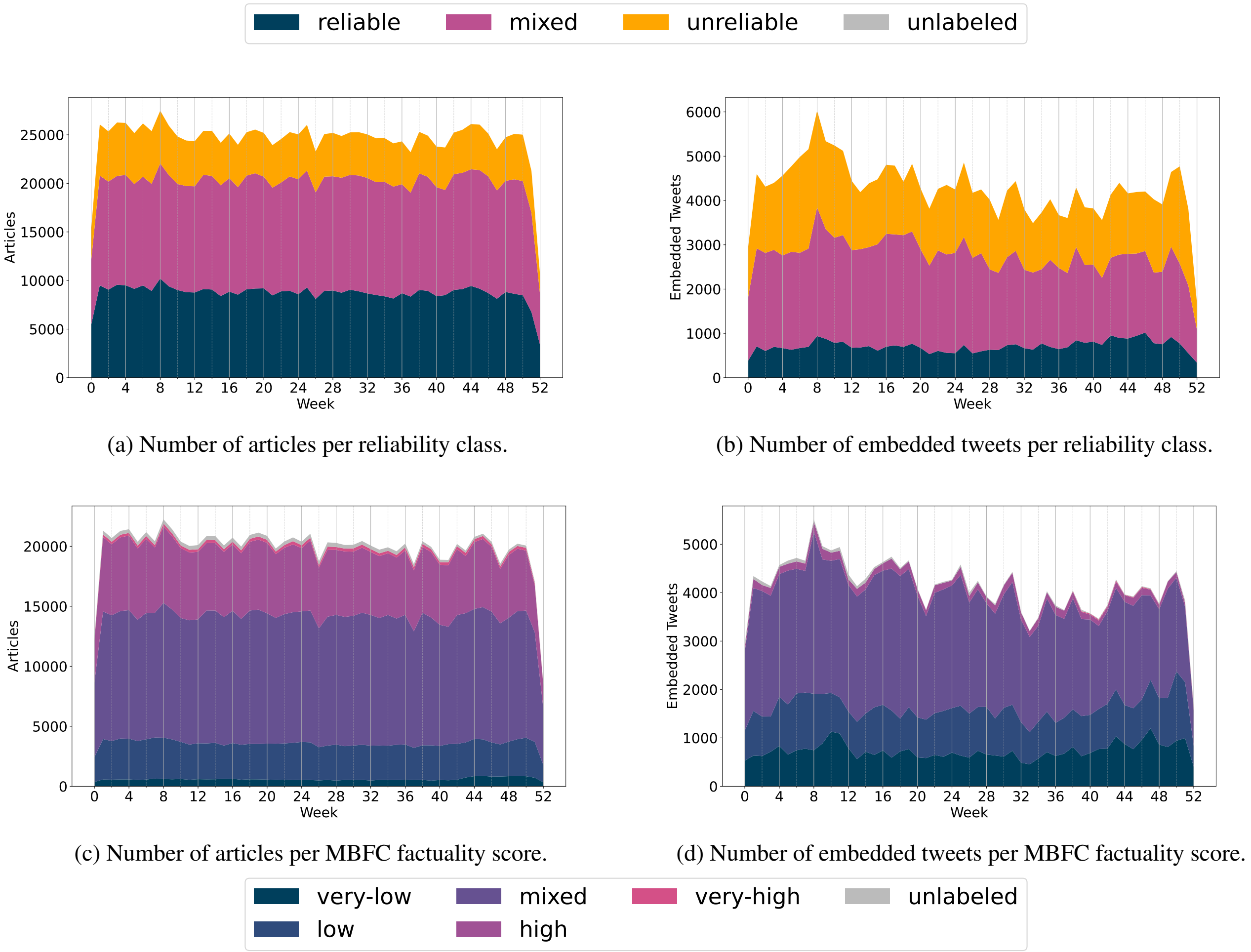
The data collection process follows what was described in (Nørregaard, Horne, and Adalı 2019). Specifically, we scraped the RSS feeds of each outlet in our outlet list twice a day starting on 01/01/2022 using the Python libraries feedparser and Goose3[1]. This list of outlets to collect was carried over from (Gruppi, Horne, and Adalı 2021). These sources come from a variety of countries (or the country of origin is not known), but are all articles are in English.

See Table 1 for details on each attribute stored during the collection process.

3.2 Embedded tweet data

In 2020, we introduced additional data on tweets embedded into news articles (Gruppi, Horne, and Adalı 2020). We again collect this data for the 2022 dataset. Specifically, we collected embedded tweets on the article page using the Goose3 library. The ID of the embedded tweet is stored in the database table tweet, along with the id of the article from which it was collected and the text of the tweet. We show this structure in Table 2.

[1] `https://github.com/grangier/python-goose`



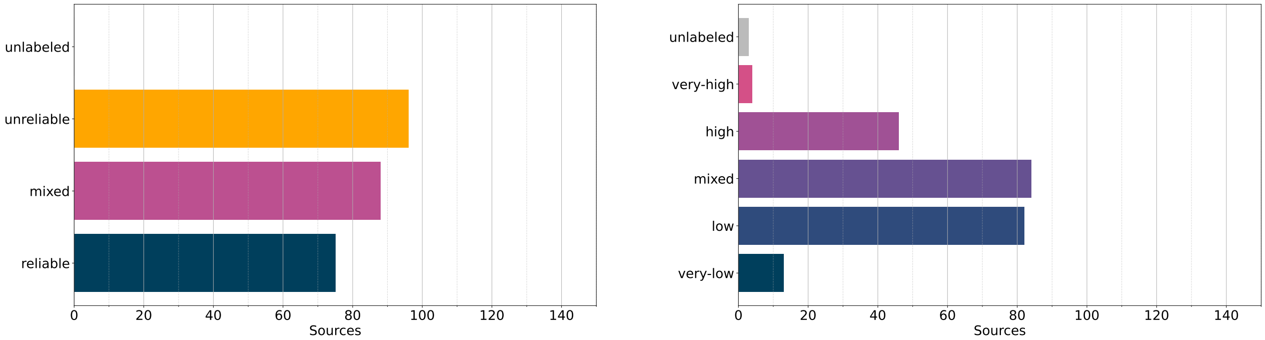
(a) Number of articles per reliability class.

(b) Number of embedded tweets per reliability class.

(c) Number of articles per MBFC factuality score.

(d) Number of embedded tweets per MBFC factuality score.

Figure 1: Number of articles (a, c) and embedded tweets (b, d) collected during each week of 2021.



(a) Number of sources per reliability class.

(b) Number of sources per MBFC factuality score.

Figure 2: Distribution of sources per reliability class (a) and factuality (b) score.

3.3 Limitations

Since the articles collected from news sources may be copyrighted, we apply a transformation to the original text so that it cannot be used for their originally intended purpose, i.e., that of being read by individuals to consume journalistic information.

Specifically, we modify the text so that it cannot properly be used for news consumption but can still be used for text analysis by periodically removing words from the text. For articles with more than 200 tokens, we replace 7 tokens with '@' every 100 tokens. For articles with fewer than 200 tokens, we replace 5 consecutive tokens with '@' every 20 tokens. This transforms the articles so that it is unlikely that a user will read NELA-GT to consume news while still keeping most of the content that is useful for analysis (approximately 7% for larger articles).

4 Format of Data

Just as in the past three versions (2019, 2020, 2021), the dataset has been released in two formats: (1) a SQLite database, (2) a JSON dictionary per news source. Details about the structure of each of these formats is below. We provide Python code to read both data formats at: `https://github.com/MELALab/nela-gt`.

The SQLite 3 database schema consists of two tables: `newsdata` and `tweet`. The `newsdata` table contains, in each row, data about an article. Column **id** is set as primary key to avoid duplicated entries on the database. We normalized source names by converting them to lower case, and removing spaces, punctuation, and hyphens. For example, the source *The New York Times* appears as *thenewyorktimes*, Tables 1 and 2 give information about data columns.

4.1 JSON Format

We also provide the dataset in JSON format. Specifically, each source has one JSON file containing the list of all of its articles. The fields follow the same structure of the database columns (Tables 1 and 2).

4.2 Ground Truth Data Format

We include multiple types of source-level veracity labels. In `NELA-GT-2022`, we collect source-level labels from Media Bias/Fact Check (MBFC) that contain the following dimensions of veracity:

1. Media Bias Fact Check factuality score - on a scale from 0 to 5 (low to high credibility).
2. Media Bias Fact Check Conspiracy/Pseudoscience and questionable sources - low credibility if a source belongs to these categories.

In addition, we create an aggregated version of the factuality scores, broken down into three classes: *reliable*, *mixed*, and *unreliable*.

Due to the limited availability of veracity labels from other platforms, we choose to only collect labels from MBFC. However, we encourage researchers to use and compare veracity labels from multiple resources when possible. This is particularly important when testing machine learning models. For an overview of the impact of ground truth labels on news studies, please see (Bozarth, Saraf, and Budak 2020). Furthermore, we strongly encourage machine learning researchers to test news veracity models using robust evaluation frameworks, such as those discussed in (Bozarth and Budak 2020) and (Horne, Nørregaard, and Adali 2019).

5 Use Cases

5.1 Analysis of news coverage during events

One of the primary goals in the yearly-release of the NELA-GT datasets is to provide updated coverage of current events. To this end, we provide two example subsets of the database for two events during 2022: the Russo-Ukrainian War and the overturning of Roe v. Wade.

Russo-Ukrainian War In February 2022, Russia launched an invasion of Ukraine, connecting back to conflict starting in 2014. This event came to many as a surprise, as Russian officials repeatedly denied plans to attack Ukraine. his international event had many ramifications across the globe, including gas pipeline disputes and global disinformation campaigns about the war[2]. This event and its effects have been widely covered in media, including fringe, conspiracy media.

Overturning of Roe v. Wade Another major event in 2022 was the overturning of Roe v. Wade in the United States, which guaranteed the right to abortion for nearly 50 years. With Roe v. Wade overturned, federal standards on abortion access were removed, allowing many states to ban abortions and forcing many women's clinics to close[3]. Given the partisan divide on abortion rights in the United States, news coverage of this event covers a range of positions across political lines.

Figure 3a and 3b show the number of article related to these events over time in the dataset.

5.2 Embedded Tweets

By providing the tweets embedded in news articles, the `NELA-GT-2022` dataset can be useful furthering in studies of political communications and hybrid media systems. Notably, very few studies have addressed low-veracity news sources role in these hybrid systems, which this dataset can aid (Gruppi et al. 2021).

5.3 Long-Term Use Cases

Our primary goal with the continued release of the NELA-GT datasets is to support long-term news research. When combining all of the NELA datasets (both the NELA-GT datasets (Nørregaard, Horne, and Adalı 2019; Gruppi, Horne, and Adalı 2020; 2021; 2022) and the original NELA2017 dataset (Horne, Khedr, and Adalı 2018)), we provide over 6.1M news articles across 5.5 years. There are multiple research avenues that this data, both in part and as a whole, supports:

[2] `https://en.wikipedia.org/wiki/Russo-Ukrainian_War`

[3] `tinyurl.com/3v5h96sr`

Column	Type	Description
id	text (primary key)	Article identifier.
date	text	Publication date string in YYYY-MM-DD format.
source	text	Name of the source from which the article was collected.
title	text	Headline of the article.
content	text	Body text of the article.
author	text	Author of the article (if available).
published	text	Publication date time string as provided by source (inconsistent formatting).
published_utc	integer	Publication time as unix time stamp.
collection_utc	integer	Collection time as unix time stamp.
url	Text	URL of the article.

Table 1: Structure of NELA-GT-2022 article data. For the database format, column **id** is the primary key of table `newsdata`.

Column	Type	Description
id	text (primary key)	Tweet id.
article_id	text (foreign key)	Id of the article in which the embedded tweet was observed.
embedded_tweet	text	ID/URL of the embedded tweet.

Table 2: Structure of NELA-GT-2022 embedded tweets. For the database format, column **id** is the primary key of table `tweet`.

- Exploring event-driven dynamics of and narratives in news media: Analyses of narrative themes before, during, and after major events continues to be a useful methodology in interdisciplinary media studies. This dataset supports these works by maintaining consistent data collection across events.
- Robust machine learning: This dataset allows for continued work in automated news veracity detection, particularly in robustness checks of current work. These robustness checks include examining prediction accuracy over time, over events, and over mixed veracity labels. We again encourage machine learning researchers to test news veracity models using robust evaluation frameworks, such as those discussed in (Bozarth and Budak 2020), and to use multiple datasets when possible.
- Examining media manipulation: Using the veracity labels in this dataset, research can examine tactics used by hyper-partisan news outlets. Additionally, with knowledge of media manipulation campaigns, such as those discussed in the Media Manipulation Casebook[4], researchers can examine how media manipulation is propagated through malicious news outlets. While there has been a substantial focus on "fake news" detection methods by researchers, there still is room to better understand media manipulation and disinformation campaigns.

6 Conclusion

In this paper, we describe the `NELA-GT-2022`, a dataset of news articles from sources of varying veracity. The RSS feeds from the sources were scraped twice a day on every day of 2022, resulting in a set with 1,778,361 articles from 361 outlets. The dataset includes the source factuality labels from Media Bias Fact Check and tweets that were embedded in the collected news articles. We provide two event-based subsets of the dataset for the study of news coverage and messaging around the war in Ukraine and the overturning of Roe v. Wade. These subsets were generated from the original dataset by filtering articles based on keyword matching.

The dataset and additional documentation can be found at: . Example code for data extraction can be found at: `https://github.com/MELALab/nela-gt`.

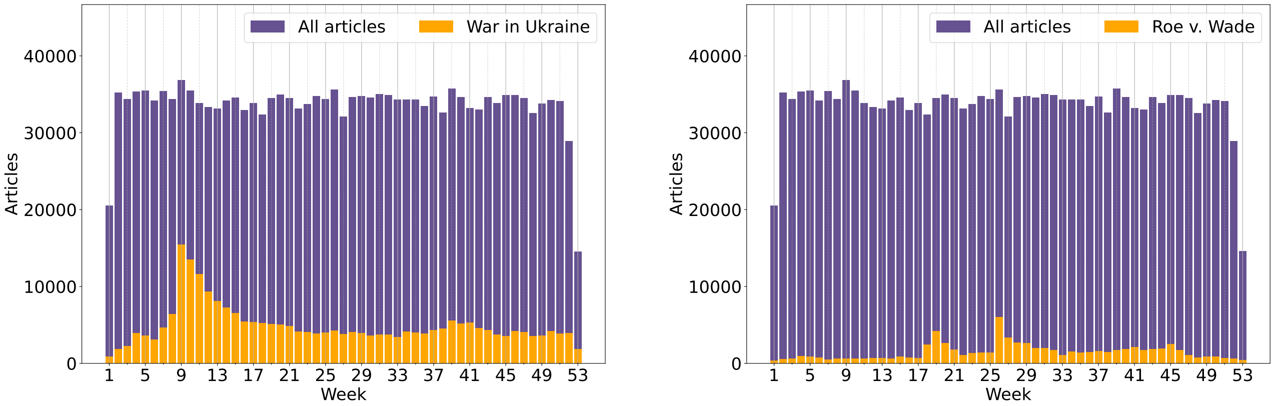
Russo-Ukrainian War keywords	Roe v. Wade keywords
Ukraine	roe
Russia	wade
Ukraine war	abortion
Ukraine and Russia	abortions
Ukrainian service members	pro-life
Ukrainian servicemen	pro-choice
Russian soldiers	overturned
Russian troops	abortion ban
Russian forces	anti-abortion
Russian-backed forces	planned parenthood
Ukrainian military	
Donetsk	
Luhansk	
Donbas	
Lyman	
Lysychansk	
Bakhmut	

Table 3: Keywords used to make the event-based data subsets for the War in Ukraine and Inflation. The full keyword lists are provided with the dataset.

References

Bode, L.; Budak, C.; Ladd, J. M.; Newport, F.; Pasek, J.; Singh, L. O.; Soroka, S. N.; and Traugott, M. W. 2020. *Words that matter: How the news and social media shaped*

[4] `https://mediamanipulation.org/`



(a) Number articles related to the Russo-Ukrainian War in comparison to all articles collected in NELA-GT-2022 in each week of 2022.

(b) Number of articles related to overturning of the Roe v. Wade case in comparison to all articles collected in NELA-GT-2022 in each week of 2022.

.

Figure 3: Number of articles related to (a) Russo-Ukrainian War, and number of articles related to (b) the overturning of the Roe v. Wade case as a fraction of the total number of articles in each week of 2022. Articles are found using a set of keywords shown in 3.

the 2016 Presidential campaign. Brookings Institution Press.

Bozarth, L., and Budak, C. 2020. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 60–71.

Bozarth, L.; Saraf, A.; and Budak, C. 2020. Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 us presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 48–59.

Deacon, D. 2007. Yesterday's papers and today's technology: Digital newspaper archives and 'push button'content analysis. *European journal of communication* 22(1):5–25.

Golbeck, J.; Mauriello, M.; Auxier, B.; Bhanushali, K. H.; Bonk, C.; Bouzaghrane, M. A.; Buntain, C.; Chanduka, R.; Cheakalos, P.; Everett, J. B.; et al. 2018. Fake news vs satire: A dataset and analysis. In *WebSci*, 17–21.

Gruppi, M.; Adalı, S.; Salemi, M.; and Horne, B. D. 2021. From tweeting about news to creating news around tweets: Characterizing tweets embedded in news articles.

Gruppi, M.; Horne, B. D.; and Adalı, S. 2020. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2003.08444*.

Gruppi, M.; Horne, B. D.; and Adalı, S. 2021. Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567*.

Gruppi, M.; Horne, B. D.; and Adalı, S. 2022. Nela-gt-2021: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2203.05659*.

Horne, B. D.; Khedr, S.; and Adalı, S. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *ICWSM*, volume 12, 518–527. AAAI.

Horne, B. D.; Nørregaard, J.; and Adali, S. 2019. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11(1):1–23.

Nørregaard, J.; Horne, B. D.; and Adalı, S. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *ICWSM*, volume 13, 630–638. AAAI.

Roberts, H.; Bhargava, R.; Valiukas, L.; Jen, D.; Malik, M. M.; Bishop, C.; Ndulue, E.; Dave, A.; Clark, J.; Etling, B.; et al. 2021. Media cloud: Massive open source collection of global news on the open web. *arXiv preprint arXiv:2104.03702*.

Salem, F. K. A.; Al Feel, R.; Elbassuoni, S.; Jaber, M.; and Farah, M. 2019. Fa-kes: a fake news dataset around the syrian war. In *ICWSM*, volume 13, 573–582.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.