Original Paper

Data Exploration and Classification of News Article Reliability: Deep Learning Study

Kevin Zhan[1*]; Yutong Li[1*], BSc; Rafay Osmani[2]; Xiaoyu Wang[3]; Bo Cao[1], PhD

[1]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada

[2]Department of Cell Biology, University of Alberta, Edmonton, AB, Canada

[3]Department of Computing Science, University of Alberta, Edmonton, AB, Canada

[*]these authors contributed equally

Corresponding Author:
Yutong Li, BSc
Department of Psychiatry
University of Alberta
4-142 KATZ Group Centre for Pharmacy and Health Research
87 Avenue and 114 Street
Edmonton, AB, T6G 2E1
Canada
Phone: 1 403 926 6628
Email: yutong5@ualberta.ca

Abstract

Background: During the ongoing COVID-19 pandemic, we are being exposed to large amounts of information each day. This "infodemic" is defined by the World Health Organization as the mass spread of misleading or false information during a pandemic. This spread of misinformation during the infodemic ultimately leads to misunderstandings of public health orders or direct opposition against public policies. Although there have been efforts to combat misinformation spread, current manual fact-checking methods are insufficient to combat the infodemic.

Objective: We propose the use of natural language processing (NLP) and machine learning (ML) techniques to build a model that can be used to identify unreliable news articles online.

Methods: First, we preprocessed the ReCOVery data set to obtain 2029 English news articles tagged with COVID-19 keywords from January to May 2020, which are labeled as reliable or unreliable. Data exploration was conducted to determine major differences between reliable and unreliable articles. We built an ensemble deep learning model using the body text, as well as features, such as sentiment, Empath-derived lexical categories, and readability, to classify the reliability.

Results: We found that reliable news articles have a higher proportion of neutral sentiment, while unreliable articles have a higher proportion of negative sentiment. Additionally, our analysis demonstrated that reliable articles are easier to read than unreliable articles, in addition to having different lexical categories and keywords. Our new model was evaluated to achieve the following performance metrics: 0.906 area under the curve (AUC), 0.835 specificity, and 0.945 sensitivity. These values are above the baseline performance of the original ReCOVery model.

Conclusions: This paper identified novel differences between reliable and unreliable news articles; moreover, the model was trained using state-of-the-art deep learning techniques. We aim to be able to use our findings to help researchers and the public audience more easily identify false information and unreliable media in their everyday lives.

(JMIR Infodemiology 2022;2(2):e38839) doi: 10.2196/38839

KEYWORDS

COVID-19; deep learning; news article reliability; false information; infodemic; ensemble model

Introduction

The onset of the COVID-19 pandemic has given the world more to battle. The world has faced a barrage of false information during the "infodemic," which is defined as the spread of a large amount of information that includes misleading or false information during a pandemic [1,2]. Due to quarantine and increased restrictions, information is trafficked to the public via social media and news sources; consequently, false information propagates at a larger scale and faster rate. Despite available public health guidelines, there is still a large presence of false and misleading information online, comprising around 20% of articles on major social media sites, such as Twitter [3]. Although the proportion of shared false information is less than evidence-informed guidelines, false information spreads at a faster rate because it contains inflammatory information [4,5]. Furthermore, infodemic management is an important aspect in maintaining public trust in scientific guidance [1]. Hence, we need to construct methods to deter the spread of false information online and identify potential sources of false news.

The abundance of fake or false news online can be instances of misinformation or disinformation and often lacks the reliability and credibility in content [6-8]. Disinformation is defined as the intentional spread of false information, while misinformation is the negligent sharing of false information [6]. Hereafter, we will not differentiate between disinformation and misinformation, as we will refer to them together as false information. False news can be categorized into 6 groups: propaganda, advertisement, manipulation, satire, parody, and fabrication [6]. Although news organizations and social media companies have implemented measures to flag and delete false news, the rate of manual false news detection is not fast enough to compete with its rapid spread through social media [9,10]. Approximately 62% of US adults obtain news from social media sites; thus, faster fact checking is critical to ensure false information spread is reduced [11]. As such, the spread of false news has resulted in public confusion, potentially associated with the antimask and vaccine rhetoric [10]. Presently, one of the most common methods to detect false news online is through human-curated fact-checking websites, such as Snopes, to flag false information [12]. Although this method may be accurate, it is inefficient due to the large amount of false news generated during the COVID-19 pandemic [10]. Thus, automatic news article reliability detection is needed.

Current false news detection using machine learning (ML) on social media has been researched extensively. Various textual features from news pages are used to predict reliability of the articles. The use of multiple features to predict the presence of false information is a common theme within current false information detection studies. The use of multiple features can improve the performance of an ML model. For example, Reis et al [13] used textual features (eg, semantic and lexical features) and news source features (eg, credibility of the news organization) as inputs for the ML model. Using traditional classifiers, such as random forest and extreme gradient boosting (XGBoost), a performance of 0.85 and 0.86 area under the curve (AUC) was achieved, respectively [13]. Elhadad et al [14] used a voting ensemble method, in addition to feature engineering, for sentiment and part-of-speech tagging. Singhania et al [15] created a 3-level HAN model using input from words, sentences, and the headline level of a news article. Similar studies have proposed that other lexical features, such as n-grams, term frequency–inverse document frequency (TF-IDF), and probabilistic context-free grammar (PCFG) have also been used as features for misinformation prediction using deep learning [16]. Accordingly, feature engineering provides higher performance metrics as well as improved interpretability. These features allow the model to focus on the important elements, which allows for reliability prediction, especially in news articles, despite high heterogeneity and noise between samples. To build on what other false information research has found, as well as to identify important new factors that contribute to false information detection, we created a final ensemble model using the ReCOVery data set [17].

Ensemble methods were implemented to further improve the performance of misinformation detection within news articles. Ensemble model usage can benefit model performance by improving the ability to generalize to data on which the model has not been trained [18]. Kumar et al [19] demonstrated improvement in performance after the use of an ensemble model, where the use of an ensemble deep learning model with a convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) was able to achieve higher performance than a CNN or long short-term memory (LSTM) model alone, with a performance of 88.78% accuracy versus 73.29% and 80.62% for the CNN and LSTM, respectively. Due to the size of news articles, a bidirectional gated recurrent unit (BiGRU) was selected as the first model in the ensemble [20]. This model is a type of recurrent neural network (RNN) that functions well on sequential text data. A BiGRU solves the vanishing gradient problem, where the model trains on long news articles and “forgets” information from the start of the articles. This model is made of many neurons or cells, each with an update gate to control what new information is added at each word and a reset gate to control how much old information is retained. A BiGRU’s bidirectional nature allows it to process each sample from the beginning and end of the article. Compared to other state-of-the-art natural language processing (NLP) models, such as LSTM, a gated recurrent unit (GRU) has lower parameters, making it quicker to train and use [21,22]. A quicker model is important as a large number of news articles are released each day; thus, a model for false information detection needs to be both accurate and fast in order to keep up with the number of new articles. XGBoost is another model included within our ensemble model. One strength of XGboost is its exceptional ability at learning from tabular data [23,24]. As a gradient boosted tree model, it is faster than a neural network and works better on the low-dimensionality output from the first model following feature extraction. Furthermore, XGBoost has been shown to outperform deep learning models for tabular data as the hyperparameter search is shorter [24]. Additionally, XGBoost combined with deep learning models in an ensemble model yields better results than an ensemble model with multiple deep learning models or classical ML models [24].

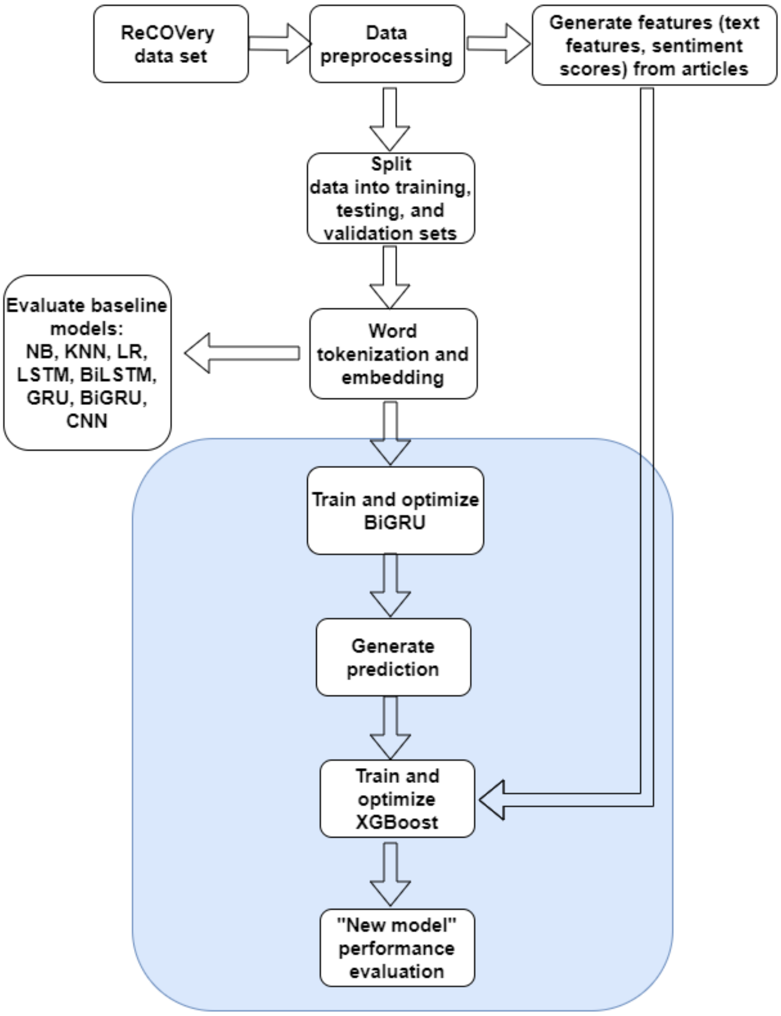
This study aims to provide a potential solution to the multifaceted false information problem through an ensemble deep learning model to classify the reliability of news articles using the ReCOVery data set. We hypothesize that sentiment, readability, lexical categories, and other text characteristics in news articles can be used together as inputs for news reliability classification improvement. We also explore differences in the sentiment or tone of reliable and unreliable information, which can be used to classify the reliability of the text. The outcome of our study may advance news reliability classification and help researchers and the public identify unreliable news articles in their everyday lives.

Methods

Workflow

First, data preprocessing was completed using the ReCOVery data set, which included removing stop words, links and Universal Resource Locators (URLs), and duplicate articles (Figure 1). Conversion of abbreviations and numbers to words was also completed within the preprocessing step. Following the preprocessing of the data, we performed feature engineering to create readability and sentiment scores, as well as extract lexical categories from the text (Figure 1). The preprocessed data were split into training, validation, and testing sets. Word tokenization and embedding were performed on the training and validation sets. Once tokenization and embedding were completed, 9 different ML models were trained and evaluated on the validation set to determine the best-performing model. We refer to naive Bayes (NB), K-nearest neighbors (KNNs), and logistic regression (LR) as traditional ML models as they are not deep learning models. The best-performing model was the ensemble model containing a bidirectional GRU and XGBoost ensemble "new model," as highlighted in blue in Figure 1.

Figure 1. Details of workflow for data exploration and "new model" construction (highlighted in blue). CNN: convolutional neural network; BiGRU: bidirectional gated recurrent unit; BiLSTM: bidirectional long short-term memory; GRU: gated recurrent unit; KNN: K-nearest neighbor; LR: logistic regression; LSTM: long short-term memory; NB: naive Bayes; XGBoost: extreme gradient boosting.


ReCOVery data set
Data preprocessing
Generate features (text features, sentiment scores) from articles
Split data into training, testing, and validation sets
Evaluate baseline models: NB, KNN, LR, LSTM, BiLSTM, GRU, BiGRU, CNN
Word tokenization and embedding
Train and optimize BiGRU
Generate prediction
Train and optimize XGBoost
"New model" performance evaluation

Data Description

The ReCOVery data set was our main source of data for news articles connected to Twitter posts [17]. It focuses on the reliability of news articles from a wide array of news sources and contains 2029 articles from ~2000 different news outlets from different countries (filtered from January to May 2020) that are related to COVID-19 news [17]. Each article was labeled as either 0 for unreliable or 1 as reliable according to the NewsGuard score [17]. The NewsGuard score was developed by journalists to label the reliability of an online article. Using a scale of 0-100, the NewGuard gives points to articles that accomplish credible and transparent news practices. Online articles with a score above 60 are labeled with a "green" rating as reliable sources, and scores below 60 are labeled with a "red" rating as unreliable sources [17,25]. In addition to the NewsGuard score, ReCOVery uses Media Bias/Fact Check, which checks the correctness of news sources according to the article subjectivity and ranks articles from “very high” to “very low” in terms of factual reporting [17,26]. Reliable articles have a NewsGuard score higher than 90, with a “very high” or “high” rating on Media Bias/Fact Check. Unreliable articles have a NewsGuard score lower than 30, with a “mixed,” “low,” or “very low” factual rating on Media Bias/Fact Check [17]. The ReCOVery data set combined the NewsGuard and Media Bias/Fact Check scores to create the final news article reliability score.

Preprocessing

Prior to data analysis, the article text and tweet data were subjected to multiple preprocessing steps. The purpose of preprocessing was to clean the data so that the deep learning model could more efficiently detect patterns in the data. The steps taken to preprocess the article text included the removal of duplicates articles or tweets; common stop words, such as “the” and “a”; and all links and non-English characters. Lemmatization of the article text was also completed, in addition to the conversion of acronyms to full terms.

Preprocessing was conducted using Python libraries, such as Pandas and Natural Language Toolkit [27,28]. A total of 1346 reliable articles and 648 unreliable articles were used for model training. Additionally, 34 articles were removed as they had less than 100 words, which limited the validity of reliability analysis. Following preprocessing, features from the news articles such as text characteristics, readability, and sentiment were extracted for analysis and to be included as input to the deep learning model.

Sentiment Analysis

Sentiment analysis was applied to the body text of reliable and unreliable articles. This was implemented through Valence Aware Dictionary and Sentiment Reasoner (VADER) and TextBlob, which are open source tools for determining predominant sentiment, polarity, and subjectivity [29,30]. The analysis relies on lexicographic analysis to map the text features of each article to different scores with regard to sentiment, polarity, and intensity. In terms of sentiment, the articles have a continuous score between 0 and 1, including both endpoints, with 1 representing that the article contains the specified sentiment as the predominant sentiment. For example, if an article has a positive sentiment of 1, this means the article contains the highest-possible positive sentiment. VADER and TextBlob were imported into Python and applied to the body text of articles within the data set. The total proportion of articles with a positive, negative, and neutral sentiment were determined through library functions within VADER and TextBlob.

Text Analysis

After preprocessing, the body text of articles was analyzed. The most common words from reliable and unreliable articles were determined. They are presented in a frequency bar graph to demonstrate the major differences between unreliable and reliable articles (Figures 2 and 3, respectively). Another feature included within the deep learning model was the text length and readability of the newspaper articles. The length of the articles was assessed using the character length of the article sentences and overall article length. Readability was assessed using 6 different readability metrics from the py-readability-metrics library: the Flesch-Kincaid grade level, Gunning fog index, Coleman-Liau index, Dale-Chall index, automated readability index (ARI), and Linsear Write index [31]. The aforementioned readability metrics are used to determine the grade level necessary to understand a written document based on the sentence length and word length [32].

The Flesch-Kincaid grade level is a scale modified from the Flesch-Kincaid reading ease index that compares the ratio of words per sentence and the ratio of syllables per word [33]. The values for this scale linearly indicate the estimated US grade level of a text. For example, a grade of 10-12 would indicate that the target reader is at the high school level, whereas scores higher than 12 are graduate-level texts [33]. Similarly, the Coleman-Liau index and the ARI both assess character and word frequency to approximate the US grade level required to read a text [34]. The Gunning fog index assesses the frequency of difficult words in a text and is a linear range between 0 and 20: a score of 16-20 is at the graduate level [35]. Similarly, the Dale-Chall index evaluates the frequency of difficult words but is scaled so that a score of 9-10 represents a university graduate–level text [31,36-38]. Lastly, the Linsear Write index was developed to assess the readability of technical texts, and its score represents the years of formal US education needed to understand a text, similar to the previous indices [39].

Topic analysis was performed using Empath, a neural network–based lexicon [40]. Empath is able to determine whether a certain sentence has the lexical categories of politics, religion, contentment, and approximately 200 more categories [40]. By processing the text with Empath, we derived 194 lexical categories that were used as additional features that were concatenated with the previous text, sentiment, and readability features in the final deep learning model. The extracted lexical categories from Empath increased the amount of information the deep learning model trained on for each article and allowed for better interpretability as differences in topic frequencies could also be evaluated. For each of the lexical categories, a mean count for reliable and unreliable articles was derived, along with the *t* test and the *P* value (Table 1).

Figure 2. Number of occurrences for keywords in unreliable news articles (N=298,498 words).

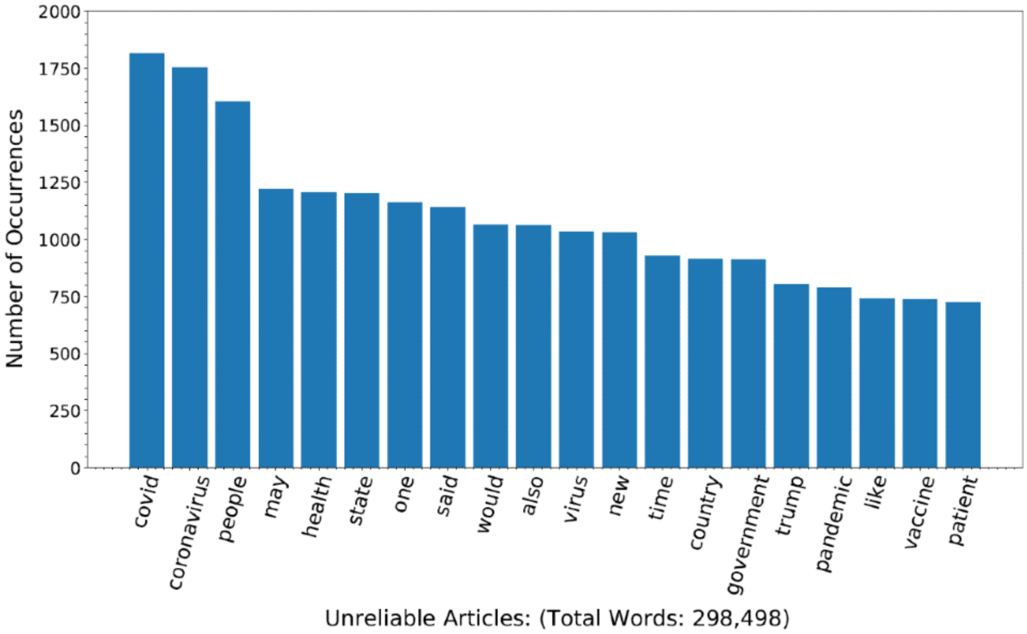


Figure 3. Number of occurrences of keywords in reliable news articles (N=662,290 words).

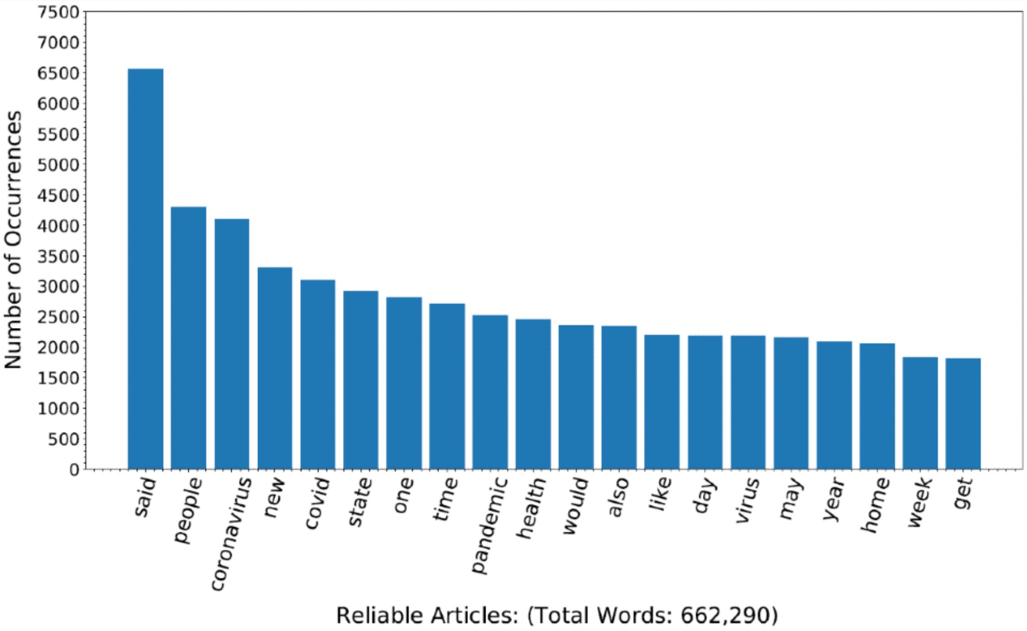


Table 1. Top 10 lexical categories from Empath (a neural network–based topic analysis tool) in reliable and unreliable news articles selected by Empath. The reliable and unreliable means is the mean counts of each lexical category being classified into reliable and unreliable news articles, respectively.

Lexical category	*t* (*df*)	*P* value	Reliable mean (SD)	Unreliable mean (SD)
magic	–7.91 (1992)	<.001	0.19 (0.60)	0.51 (1.22)
power	–7.16 (1992)	<.001	1.28 (2.20)	2.16 (3.24)
business	7.15 (1992)	<.001	8.58 (10.54)	5.31 (7.10)
work	6.89 (1992)	<.001	5.78 (8.82)	3.28 (3.89)
contentment	6.18 (1992)	<.001	0.70 (1.61)	0.29 (0.72)
office	6.14 (1992)	<.001	3.02 (4.37)	1.88 (2.60)
dispute	–6.11 (1992)	<.001	1.58 (2.48)	2.35 (2.94)
morning	5.87 (1992)	<.001	1.06 (1.87)	0.59 (1.11)
legend	–5.85 (1992)	<.001	0.34 (0.92)	0.64 (1.31)
blue collar job	5.83 (1992)	<.001	0.62 (1.75)	0.21 (0.68)

Tokenization

As ML models only accept numerical inputs, the text data must be tokenized. This process involves a word-index dictionary, where each word in the data set is converted to a numerical value or index, which corresponds to that word in the dictionary. For example, a word such as “coronavirus” might be presented to a ML model as the value 1234. As each unique word creates a unique index number, the “vocabulary” or total number of unique words in the data set can be a problem, especially if the data set is large, since words that appear once or twice in the data set generally do not contribute to the efficacy of the model. We limited the vocabulary size to 20,000 (51.73%) out of a total of 38,663 unique words from the training data. This excluded words that were used only once in the data set, because these words would not be useful to the model—Zipf’s law reaffirms that having larger vocabulary sizes gives diminishing returns as we frequently use a small proportion of their total vocabulary [41,42]. Furthermore, there are various estimates regarding the vocabulary size of an average native English speaker, with around 20,000 being a reasonable estimate for the vocabulary size [43,44]. Articles were also 0-padded to a size of 3500 words, which was the size of the longest article to ensure uniformity of the model input.

Word Embedding

Following tokenization, the data were processed using word embedding, a form of unsupervised ML. Word embedding places the data points of individual words into an embedding space with high dimensionality. Inside this embedding space, each word is represented as a vector with words that are similar to each other being located in close proximity. As such, word embedding allows hidden relationships between similar words to be quantified for ML analysis. Although a new word embedding layer could be trained and fitted on our data set, there exist pretrained word embedding models that are more efficient to use. For the article text data, we leveraged Global Vectors for Word Representation (GloVE), which is a commonly used word embedding model trained on hundreds of thousands of Wikipedia articles, which have an embedding space of 100 dimensions [45].

Machine Learning Classification

The data were randomly split into training, testing, and validation subsets for deep learning. The ratio of these subsets was 8:1:1, respectively. Of the 1994 articles, 1595 (79.99%) were in the training subset, 199 (9.98%) were in the validation subset, and 200 (10.03%) were in the testing subset. The training and validation data were used to build the model to classify between reliable and unreliable articles, while the testing data were used to evaluate the model’s performance. The splitting of the data followed by model training and evaluation were repeated 10-folds so that each article could be included in the training set. An average was taken between the performance metrics obtained from training on each fold. We evaluated the performance of multiple ML models on the data set (NB, KNNs, LR, LSTM, GRU, BiLSTM, BiGRU, and CNN) to determine the best models for reliability detection. The settings or hyperparameters were optimized either experimentally or using Gridsearch, which tests all combinations of hyperparameters for each of the aforementioned ML models.

Finally, we developed an ensemble model using a lightly trained BiGRU to generate an initial reliability prediction, which was then combined with the text features, readability, sentiment, and Empath-classified lexical categories. This was then used to train an XGBoost model with 10-fold cross-validation.

This paper uses several evaluation metrics that rely on the results from the confusion matrix. These metrics were derived from correct predictions by the model, such as true positive (TP) and true negative (TN), as well as incorrect predictions, such as false positive (FP) and false negative (FN). Accuracy is the total proportion of correct predictions, but this evaluation metric is not as effective when there is a class imbalance. Sensitivity refers to the proportion of correctly predicted unreliable articles, while specificity refers to the proportion of correctly predicted reliable articles. The AUC score shows the performance of the model at different TP and FP rates [46].

$$\text{Sensitivity (recall)} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Ethical Considerations

The data used in this paper did not need ethics approval as they were accessed through the open access ReCOVery data set GitHub, as cited in Zhou et al [17].

Results

Data Exploration

Data exploration was performed and features, such as readability, sentiment, and lexical categories, were combined with the full news article text data to train an ensemble model. An ensemble method using BiGRU and XGBoost was created using 1346 reliable articles and 648 unreliable articles.

During data exploration, we found that the average text length in terms of the average word length and sentence length was longer in unreliable articles compared to reliable articles (Table 2). The Flesch-Kincaid grade level, the Dale-Chall index, the ARI, the Coleman-Liau index, the Gunning fog index, and the Linsear Write index indicated that reliable articles are easier to read compared to unreliable articles (Table 2). From the average frequency of 194 Empath-derived lexical categories, 110 (56.7%) were significantly different between reliable and unreliable articles (Multimedia Appendix 1). Most frequent words in unreliable and reliable articles were also visualized (Figures 2 and 3, respectively). Unreliable articles had higher rates of negative sentiment, while reliable articles had higher rates of neutral sentiment (Table 3). Performance metrics of various trained ML models as well as the new ensemble model were determined (Table 3).

Table 2. Text length and readability metrics for reliable (N=1346) and unreliable (N=648) online news articles. The text length was expressed as the average sentence length and word length. Readability was expressed using the Flesch-Kincaid grade level, the Dale-Chall readability index, the ARI[a], the Coleman-Liau index, the Gunning fog index, and the Linsear Write index.

Metrics	Reliable mean (SD)	Unreliable mean (SD)	*t* (*df*)	*P* value
Average word length (characters)	6.14 (0.27)	6.32 (1.66)	–3.93 (1992)	<.001
Average sentence length (words)	23.67 (5.17)	26.38 (7.06)	–9.70 (1992)	<.001
Flesch-Kincaid grade level	12.68 (2.63)	14.39 (3.37)	–12.38 (1992)	<.001
Gunning fog index	14.87 (2.72)	16.42 (3.33)	–11.00 (1992)	<.001
Coleman-Liau index	10.85 (1.87)	11.82 (2.46)	–9.72 (1992)	<.001
Dale-Chall index	10.21 (0.96)	10.70 (1.02)	–10.53 (1992)	<.001
ARI	13.41 (3.30)	15.43 (4.47)	–11.41 (1992)	<.001
Linsear Write index	16.42 (4.02)	18.73 (5.31)	–10.80 (1992)	<.001

[a]ARI: automated readability index.

Table 3. Comparison of sentiment polarity (0=least expression of sentiment in interest, 1=most expression of sentiment in interest) between reliable (N=1346) and unreliable (N=648) news articles in terms of sentiment of the sentences within news articles. Differences between the frequencies of sentences possessing positive, neutral, or negative sentiment were analyzed with a 2-sample independent *t* test.

Sentiment	Reliable mean (SD)	Unreliable mean (SD)	*t* (*df*)	*P* value
Negative	0.066 (0.042)	0.076 (0.039)	–5.46 (1992)	<.001
Neutral	0.850 (0.054)	0.840 (0.050)	4.37 (1992)	<.001
Positive	0.084 (0.035)	0.085 (0.035)	–0.095 (1992)	.92

Text Analysis

After removal of stop words, the most frequent words in reliable and unreliable articles were examined. The highest word frequencies for unreliable and reliable articles are illustrated in frequency bar graphs (Figures 2 and 3). Common words between reliable and unreliable news articles were COVID-19–related keywords, such as “coronavirus,” “virus,” and “pandemic.” The differences were related to political undertones, such as “Trump” and “government.” Additionally, the Empath lexicon tool was applied to the text to yield lexical categories. The average count for each lexical category was determined for reliable and unreliable text. The differences in means were then compared using *t* tests. There were a total of 194 lexical categories that significantly differed in frequency between reliable and unreliable texts (Multimedia Appendix 1 and Table 1). In Table 1, we display the top 10 lexical categories with the lowest *P* value. Categories included “magic,” “power,” “business,” “work,” “contentment,” “office,” “dispute,” “morning,” “legend,” and “blue collar job.” The lexical categories “business,” “work,” “contentment,” “office,” “morning,” and “blue collar job” had higher mean counts for the reliable articles compared to the unreliable articles. The lexical categories “magic,” “power,” “legend,” and “dispute” had lower mean counts for the reliable articles compared to the unreliable articles. In terms of text characteristics, there was a significant difference in the average sentence length between reliable and unreliable news articles, with reliable articles containing shorter sentences at 23.67 (SD 5.17) words per sentence compared to unreliable articles containing 26.38 (SD 7.06) words per sentence (Table 2). Additionally, the average word lengths were 6.14 (SD 0.27) and 6.32 (SD 1.66) for reliable and unreliable articles, respectively. In addition to text length, we also analyzed the differences in readability between reliable and unreliable articles. The readability indices used were the Flesch-Kincaid grade level, the Dale-Chall index, the ARI, the Coleman-Liau index, the Gunning fog index, and the Linsear Write index. As shown in Table 2, unreliable articles were less readable, as indicated by all 6 readability indices. Since these text features are important in differentiating between reliable and unreliable news articles, they were input into our final deep learning model.

Sentiment Analysis

Using VADER, the sentences from the articles were classified into positive, neutral, and negative sentiments. The sentiment score ranged from 0 to 1, with 1 denoting strong presentation of the sentiment of interest. For reliable articles, the means for the negative, neutral, and positive sentiments scores were 0.066 (SD 0.042), 0.850 (SD 0.054), and 0.084 (SD 0.035), respectively (Table 3). For unreliable articles, the means for the negative, neutral, and positive sentiment scores were 0.076 (SD 0.039), 0.840 (SD 0.050), and 0.084 (SD 0.035), respectively.

Machine Learning Analysis

After the newspaper article data were passed through GloVE word embedding, the text data were split 10-folds for cross-validation. The traditional ML models included LR, KNNs, and NB. The AUC values (Figure 4) were generated, in addition to sensitivity and recall values (Table 4).

Next, the deep learning models were fit to the data. Each model included the GloVE word embedding prior to training. Optimization of hyperparameters for the deep learning models was completed using GridSearchCV from the ML Python scikit-learn library. The hyperparameters optimized were batch size, epochs, dropout rate, neuron number, optimizer type, learning rate, and activation function type. Each model had varying hyperparameters that yielded the best results.

The deep learning models that were assessed were LSTM, GRU, BiLSTM, BiGRU, and CNN. Similar to traditional ML models, the AUC, specificity, and recall were determined as performance metrics (Table 4).

Lastly, an ensemble model was developed using the BiGRU and XGBoost. Our new model was first evaluated on the ReCOVery testing subset. A confusion matrix for our new model was generated, as shown in Figure 5. The AUC, specificity, and sensitivity for our new deep learning model were 0.906, 0.835, and 0.945, respectively (Table 4).

Figure 4. Receiver operating characteristic (ROC) curve and AUC scores with the corresponding color for both traditional ML models (KNN, LR,NB) and deep learning models (BiLSTM, CNN, LSTM, BiGRU, GRU, new model). AUC: area under the curve; BiGRU: bidirectional gated recurrent unit; BiLSTM: bidirectional long short-term memory; CNN: convolutional neural network; FP: false positive; GRU: gated recurrent unit; KNN: K-nearest neighbor; LR: logistic regression; LSTM: long short-term memory; ML: machine learning; NB: naive Bayes; TP: true positive.

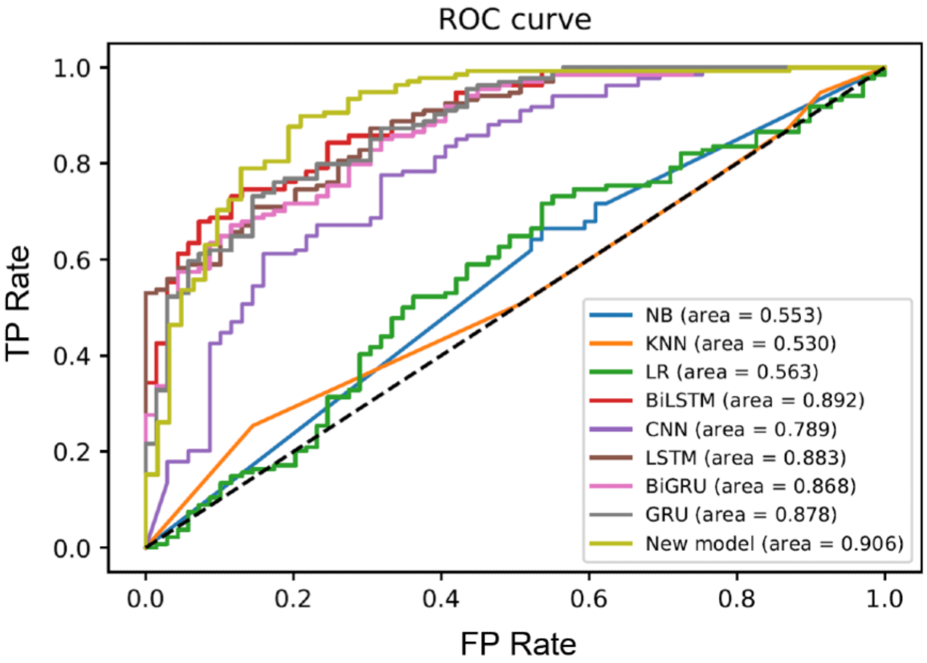
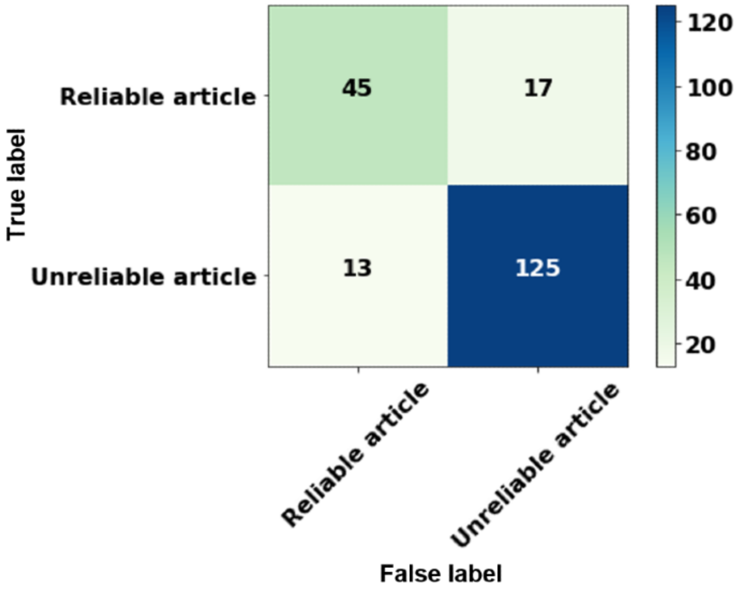


Table 4. Performance metrics for the ReCOVery validation data set for traditional ML[a] models (KNN[b], LR[c], NB[d]), and deep learning models (BiLSTM[e], CNN[f], LSTM[g], BiGRU[h], GRU[i], new model).

Model	Specificity	Sensitivity	AUC[j]
LR	0.720	0.575	0.563
KNN	0.660	0.739	0.530
NB	0.700	0.627	0.553
BiLSTM	0.810	0.925	0.892
CNN	0.792	0.851	0.789
LSTM	0.829	0.903	0.883
BiGRU	0.791	0.963	0.868
GRU	0.804	0.918	0.878
New model	0.835	0.945	0.906

[a] ML: machine learning.
[b] KNN: K-nearest neighbor.
[c] LR: logistic regression.
[d] NB: naive Bayes.
[e] BiLSTM: bidirectional long short-term memory.
[f] CNN: convolutional neural network.
[g] LSTM: long short-term memory.
[h] BiGRU: bidirectional gated recurrent unit.
[i] GRU: gated recurrent unit.
[j] AUC: area under the curve.

Figure 5. Confusion matrix for ReCOVery validation subset on trained new ensemble model with BiGRU and XGBoost. BiGRU: bidirectional gated recurrent unit; XGBoost: extreme gradient boosting.



Discussion

Principal Findings

This study demonstrates an ensemble model with BiGRU and XGBoost for text reliability classification using the ReCOVery data set with a specificity, sensitivity, and AUC of 0.835, 0.945, and 0.906, respectively [17]. Through our data analysis, we demonstrated that unreliable news articles have lower readability and higher sentence length. They also include more negative and less neutral sentiments and contain more polarizing lexical categories in comparison to reliable articles.

Data Usage

With regard to using news articles to build a classification model, an important consideration is the generalizability of the model. To ensure that the model is generalizable, the data used to train the model must be diverse in nature. A shortcoming of many deep learning misinformation detection studies is the focus on detecting misinformation from a narrow range of news sources, or locations [17,47]. Because of the homogenous nature of the data set used to train these models, many misinformation detection models are potentially less generalizable [47]. An example would be CoAID, a data set constructed from COVID-19–related news articles and social media posts from December 1, 2019, to September 1, 2020. A shortcoming of the CoAID data set would be the lower number of news sources used for the data set as 9 reliable news sources were included during the data collection process [48]. CoVerifi is a study that used the CoAID data set to create a web-based tool to check whether an online news article was credible [49]. Another notable data set is the COVID-19-FAKES data set containing 61,711 tweets with misinformation and 2,985,399 tweets without misinformation [50,51]. Silva et al [51] used the COVID-19-FAKES data set to obtain insights into predictive features for the presence of misinformation in tweets and the differential engagement in tweets with and without misinformation [51]. Hence, we used the ReCOVery data set for the diverse nature of the news articles as they range from ~2000 different news outlets from different countries [17].

Sentiment Analysis

VADER was used to evaluate sentiment at a lexicon-based level due to its high accuracy, with an *F*1 classification accuracy of 0.96 and computational economy [29]. Although VADER has become a staple in NLP for sentiment analysis, 2 key shortcomings to consider are its inability to recognize sarcasm/satire and its reduced accuracy when dealing with 3-class analyses (negative, neutral, and positive) [52].

From the distribution of articles with reliable versus unreliable news articles, it can be observed that reliable articles contain less negative sentiment in comparison to unreliable articles as they had a lower negative sentiment polarity score (Table 3). This is in line with observations of news content in the literature, as Arif et al [53] discussed how individuals searching for negative terms on the internet can lead to more biased articles. To emphasize the importance of sentiment in differentiating fake and real news, Paschen [54] concluded that the titles and body text of fake news articles contain more negative content, such as anger and disgust, compared to real news articles. Fake news is more likely to display negative sentiment to drive a specific narrative for profit, which supports our finding that there are a greater number of negative unreliable sources than neutral or positive unreliable sources.

We observed a difference between the number of neutral reliable and neutral unreliable articles, with more neutral sentiment in reliable articles in comparison to unreliable articles (Table 3). A neutral sentiment scoring for reliable data sources implies impartiality and objectivity when discussing the subject matter [55].

Many ML studies have targeted sentiment as a feature to predict misinformation in a variety of written information online because of the different sentiment valence between reliable and unreliable text due to the aforementioned reasons [56]. Because of the differing nature of sentiment between texts of differing reliability, sentiment analysis was used in the context of filtering out negative messages on social media, spam filtering, among other applications [56]. In agreement with our findings, Ajao et al [57] determined that unreliable tweets often contain more negative sentiment in comparison to reliable tweets due to how authors of unreliable tweets use negative emotions to better propagate their message. They also showed that the use of sentiment can boost support vector machine (SVM) accuracy when the sentiment is considered in addition to textual features [57]. Hence, sentiment was a feature selected for our model.

Text Analysis

The words themselves were observed to be quite similar to one another between the 2 groups because the subject matter of both reliable and unreliable sources is the same: COVID-19. Additionally, many of the most frequently occurring words are mere transitional words that are likely to be found in the majority of English literature.

Interestingly, the most frequently occurring word in reliable sources was “said” (Figure 3). This is likely due to “said” being used to quote political figures and leaders in the scientific field. The reliability of articles in this case is a consequence of the articles citing reliable sources of information. Another observable trend is the increasing number of politically charged words found in unreliable articles. Words such as “country,” government,” and “Trump” were amongst the most frequent words for unreliable sources but not for reliable articles (Figure 2). This communicates a pattern of political commentary occurring in unreliable sources [58]. We can anticipate that articles discussing political content in the context of COVID-19 are likely interested in propagating an agenda—hence, the unreliability. For example, Chen et al [59] found interplay between COVID-19 misinformation propagation and the 2020 US presidential elections with regard to mask use and mail-in ballots. Specifically, health information has been politicized to push political agendas and attack political opponents. In addition to frequently occurring words, lexical categories extracted from Empath and similar models allows us to evaluate the difference in topic frequencies between reliable and unreliable news articles [40]. The use of lexical categories extracted from Empath and similar models can increase model performance compared to using only raw text data [60-63].

Another feature we decided to explore and include in our final deep learning model is the readability and length of the news articles. Readability has been shown to be predictive of misinformation. In the study by Santos et al [64], articles from a frequent source of fake news could be differentiated using only article readability scores with an SVM algorithm with an accuracy of 92% [64]. Similarly, in a study by Zhou et al [65], various metrics were explored based on their ability to classify reliable versus unreliable news articles. It was determined using random forests that readability is among the top 5 in terms of contribution to the model, alongside sentiment [65].

Machine Learning Classification

In the original ReCOVery study, Zhou et al [17] created a baseline prediction performance for news article reliability and found that a precision of 0.721-0.836 and 0.421-0.667 can be obtained for reliable and unreliable news articles, respectively.

A recall of 0.705-0.829 and 0.441-0.667 can be obtained for reliable and unreliable news articles, respectively [17]. The features used in the baseline model ranged from text lexical categories, rhetorical structure, and visual information within news articles. Zhou et al [17] also tested the model on traditional ML models, such as SVMs, or deep learning algorithms, such as CNNs with unimodal and multimodal features. Other studies have also explored the use of the ReCOVery data set for false information classification. One such study is by Raj and Meel [66], where a novel deep learning model, the Allied Recurrent and Convolutional Neural Network (ARCNN), was created using both image and textual features within news articles to detect misinformation. The performance of the ARCNN was tested using 6 COVID-19 fake news data sets, with ReCOVery as 1 of the data sets, achieving an accuracy, precision, recall, and *F*1 score of 80.98%, 53.85%, 58.33%, and 56.00%, respectively [66]. Another study using the ReCOVery data set for model development explored the use of multiple languages for fake news detection to improve model performance [67]. Finally, Wahle et al [68] used the ReCOVery data set as 1 of 6 COVID-19 misinformation data sets to evaluate the performance of 15 transformer-based ML models to determine the generalizability of different transformer models. Differing from the aforementioned studies, we were able to demonstrate that the use of readability, text characteristics, sentiment, and lexical categories can improve upon the original ReCOVery data set baseline models [17]. Hence, we demonstrate the importance of the aforementioned text features to improve upon news article reliability prediction. Furthermore, we show that the combination of multiple inputs and consideration of the chosen model can increase ML model accuracy in the context of NLP.

In our final proposed model, the BiGRU with XGBoost and feature engineering was the best-performing model. A BiGRU is adept at capturing temporal data in long sequences, as bidirectional models can better capture the context of the text [46]. During the experimentation with these models on ReCOVery data, we found that all deep learning models outperformed the traditional ML models because deep learning models are better able to handle more complex data [46,69]. Furthermore, we chose to use the GRU algorithm, which is a variant of the recurrent neural network, in addition to the LSTM algorithm due to the increased performance on longer text compared to LSTM [21]. To further increase the performance of our model, an ensemble model was built, as combining multiple predictions can yield more accurate predictions [70].

Strengths

A strength of our investigation is that it not only had the main goal of creating a deep learning model for reliability prediction but also identified significant trends in text and sentiment for reliable and unreliable news articles. An investigation focused solely on a deep learning model has a "black box" problem where the mechanisms used by the deep learning model are not visible and are contained within its many complex hidden layers [71]. As such, a data exploration approach coupled with the deep learning model is able to better visualize and portray article reliability classification. Furthermore, our paper examined news articles, which had the advantage of being more normalized in text compared to tweets and social media as, each article was written with a professional approach. As such, less data were removed during preprocessing due to grammatical or spelling errors. Using news articles as data also avoided the problem of low hydration that Twitter misinformation data sets suffer from when tweets are removed by Twitter.

Limitations and Future Directions

There are a number of ways our project could be further refined. First, expanding the number of total available data would be valuable as there are nearly twice as much data for reliable sources as unreliable. Furthermore, being able to web-scrape Facebook postings and Reddit threads would allow us to expand our scope of access and evaluate other high-traffic sources of information. Incorporating clustering models would also increase the specificity of our search and create a more accurate model that can consider what aspect of COVID-19 is being discussed when determining reliability. Due to the high accuracy of our model, as shown by the results, our model can be commercialized as a web app that allows users to assess, to a high degree of confidence, the reliability of the article they are reading. Moreover, it can also be used to determine the sentiment scoring of an article to determine whether they want to engage in that specific literature.

Although this model specifically identifies COVID-19–related information, it could also be trained for other types of misinformation. As discussed previously, most current methods to combat misinformation online are through the use of human-moderated fact-checking websites. Examples include Twitter's Birdwatch program, where independent users can flag posts they deem untrustworthy [72]. Other methods used include Facebook's fact-checking service, which manually labels posts or websites containing misinformation as untrustworthy and removes them from public view [73]. Furthermore, warnings are placed below posts containing COVID-19 information to warn readers regarding potential misinformation contained within posts [73]. Even though there are numerous instances of fact checking, the major issue that arises is the inefficiency in manual fact checking [74]. Hence, new fact-checking methods aim toward automating the fact-checking process. The first example of a fact-checking website is the Bot Sentinel automated Twitter fact-checking software, which can be installed by users to monitor spam accounts [75]. Bot Sentinel uses ML technology to classify posts or profiles as reliable or unreliable to an accuracy of 95% [75].

Conclusion

In conclusion, we demonstrated that readability, sentiment, text characteristics, and lexical categories are important in differentiating between reliable and unreliable news articles, as it was shown that unreliable articles are less readable, have more negative sentiment, and have more political lexical categories. The aforementioned features were used to achieve above-the-baseline performance within the original ReCOVery data set, with a specificity, sensitivity, and AUC of 0.835, 0.945, and 0.906, respectively, using our new ensemble deep learning model. Hence, the application of readability, sentiment, and lexical categories using our new model can help determine the dependability of news articles and better improve upon pre-existing models that do not use these features.

COVID-19 has brought to light the importance of developing an automated reliability assessor for news articles, as human-moderated fact-checking methods may be inefficient. Because readability, sentiment, and lexical categories can be used to improve upon pre-existing reliability classification models, we show that automated reliability detection may be an alternate way to determine new article reliability in the future, which will help news readers identify articles containing potentially unreliable information.

Acknowledgments

We would like to thank the science, technology, engineering, and mathematics (STEM) fellowship team for organizing the National Undergraduate Big Data Challenge 2021 and JMIR for sponsoring this publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Mean (SDs) scores for Empath categories of reliable and unreliable news articles.
[XLSX File (Microsoft Excel File), 31 KB-Multimedia Appendix 1]

References

1. Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. J Med Internet Res 2020 Jun 29;22(6):e21820 [FREE Full text] [doi: 10.2196/21820] [Medline: 32589589]
2. World Health Organization. Infodemic. 2022. URL: https://www.who.int/health-topics/infodemic [accessed 2022-06-15]
3. Obiała J, Obiała K, Mańczak M, Owoc J, Olszewski R. COVID-19 misinformation: accuracy of articles about coronavirus prevention mostly shared on social media. Health Policy Technol 2021 Mar 10;10(1):182-186 [FREE Full text] [doi: 10.1016/j.hlpt.2020.10.007] [Medline: 33163352]
4. Figueira Á, Oliveira L. The current state of fake news: challenges and opportunities. Procedia Comput Sci 2017;121:817-825. [doi: 10.1016/j.procs.2017.11.106]
5. Sylvia Chou W, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health 2020 Oct;110(S3):S273-S275. [doi: 10.2105/ajph.2020.305905]
6. Tandoc E, Lim Z, Ling R. Defining "fake news". Digit J 2018;6(2):137-153. [doi: 10.1002/9781119011071.iemp0300]
7. Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, et al. The science of fake news. Science 2018 Mar 09;359(6380):1094-1096. [doi: 10.1126/science.aao2998] [Medline: 29590025]
8. Tandoc EC. The facts of fake news: a research review. Sociol Compass 2019 Jul 25;13(9):e12724. [doi: 10.1111/soc4.12724]
9. de Oliveira NR, Pisa PS, Lopez MA, de Medeiros DSV, Mattos DMF. Identifying fake news on social networks based on natural language processing: trends and challenges. Information 2021 Jan 18;12(1):38. [doi: 10.3390/info12010038]
10. Torabi Asr F, Taboada M. Big data and quality data for fake news and misinformation detection. Big Data Soc 2019 May 23;6(1):205395171984331. [doi: 10.1177/2053951719843310]
11. Clayton K, Blair S, Busam JA, Forstner S, Glance J, Green G, et al. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Polit Behav 2019 Feb 11;42(4):1073-1095. [doi: 10.1007/s11109-019-09533-0]
12. Krause NM, Freiling I, Beets B, Brossard D. Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. J Risk Res 2020 Apr 22;23(7-8):1052-1059. [doi: 10.1080/13669877.2020.1756385]
13. Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F. Supervised learning for fake news detection. IEEE Intell Syst 2019 Mar;34(2):76-81. [doi: 10.1109/mis.2019.2899143]
14. Elhadad M, Li K, Gebali F. An ensemble deep learning technique to detect COVID-19 misleading information. In: Advances in Networked-Based Information Systems. Manhattan, NY: Springer International; 2020 Aug 20 Presented at: International Conference on Network-Based Information Systems; 2020; Victoria, Canada p. 163-175. [doi: 10.1007/978-3-030-57811-4_16]
15. Singhania S, Fernandez N, Rao S. 3HAN: a deep neural network for fake news detection. 2017 Presented at: 24th International Conference on Neural Information Processing (ICONIP 2017); 2017; Guangzhou, China. [doi: 10.1007/978-3-319-70096-0_59]
16. Khan JY, Khondaker MTI, Afroz S, Uddin G, Iqbal A. A benchmark study of machine learning models for online fake news detection. Mach Learn Appl 2021 Jun;4:100032. [doi: 10.1016/j.mlwa.2021.100032]
17. Zhou X, Mulay A, Ferrara E, Zafarani R. ReCOVery: a multimodal repository for COVID-19 news credibility research. 2020 Presented at: CIKM '20: 29th ACM International Conference on Information & Knowledge Management; October 19-23, 2020; Virtual Event Ireland p. 3205-3212. [doi: 10.1145/3340531.3412880]
18. Ganaie M, Hu M, Malik A, Tanveer M, Suganthan P. Ensemble deep learning: a review. arXiv Preprint posted online April 6, 2021 [FREE Full text] [doi: 10.1016/j.engappai.2022.105151]

19. Kumar S, Asthana R, Upadhyay S, Upreti N, Akbar M. Fake news detection using deep learning models: a novel approach. Trans Emerging Tel Tech 2019 Nov 05;31(2):e3767. [doi: 10.1002/ett.3767]
20. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv Preprint posted online December 11, 2014. [doi: 10.48550/arXiv.1412.3555]
21. Yang S, Yu X, Zhou Y. LSTM and GRU neural network performance comparison study: taking yelp review dataset as an example. 2020 Presented at: 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI); June 12-14, 2020; Shanghai, China. [doi: 10.1109/iwecai50956.2020.00027]
22. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
24. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion 2022 May;81:84-90. [doi: 10.1016/j.inffus.2021.11.011]
25. NewsGuard Technologies. Rating Process and Criteria. 2022. URL: https://www.newsguardtech.com/ratings/rating-process-criteria/ [accessed 2022-06-14]
26. Zandt D. About Media Bias / Fact Check. URL: https://mediabiasfactcheck.com/about/ [accessed 2022-06-14]
27. Loper E, Bird S. NLTK: the natural language toolkit. arXiv Preprint posted online May 17, 2002 [FREE Full text] [doi: 10.3115/1118108.1118117]
28. McKinney W. Data structures for statistical computing in Python. 2010 Presented at: 9th Python in Science Conference (SciPy 2010); June 28-July 3, 2010; Austin, TX p. 56-61. [doi: 10.25080/majora-92bf1922-00a]
29. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: 8th International AAAI Conference on Weblogs and Social Media; June 1-4, 2014; Ann Arbor, MI URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550
30. Loria S. TextBlob: Simplified Text Processing. URL: https://textblob.readthedocs.io/en/dev/ [accessed 2022-09-13]
31. DiMAscio C. py-readability-metrics. URL: https://github.com/cdimascio/py-readability-metrics [accessed 2022-03-01]
32. Si L, Callan J. A statistical model for scientific readability. 2001 Presented at: CIKM01: 10th International Conference on Information and Knowledge Management; October 5-10, 2001; Atlanta, GA p. 574-576. [doi: 10.1145/502585.502695]
33. Hoke B. Comparison of Recreational Reading Books Levels Using the Fry Readability Graph and the Flesch-Kincaid Grade Level. 1999. URL: https://eric.ed.gov/?id=ED428333 [accessed 2022-06-16]
34. Karmakar S, Zhu Y. Visualizing multiple text readability indexes. 2010 Presented at: International Conference on Education and Management Technology (ICEMT 2010); November 2-4, 2010; Cairo, Egypt. [doi: 10.1109/icemt.2010.5657684]
35. Gunning R. The fog index after twenty years. J Bus Commun 2016 Sep 16;6(2):3-13. [doi: 10.1177/002194366900600202]
36. Burke V, Greenberg D. Determining readability: how to select and apply easy-to-use readability formulas to assess the difficulty of adult literacy materials. Adult Basic Educ Lit J 2010;4(1):34-42.
37. Dale E, Chall J. A formula for predicting readability. Educ Res Bull 1948;27(1):11-28.
38. Scott B. The New Dale-Chall Readability Formula: A Vocabulary-Based Readability Formula. 2003. URL: https://www.readabilityformulas.com/new-dale-chall-readability-formula.php [accessed 2022-06-17]
39. Scott B. How to Use the Linsear Write Readability Formula to Grade Your Text. 2003. URL: https://readabilityformulas.com/linsear-write-readability-formula.php [accessed 2022-06-17]
40. Fast E, Chen B, Bernstein M. Empath: understanding topic signals in large-scale text. 2016 Presented at: ACM CHI 2016: 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, CA p. 4647-4657. [doi: 10.1145/2858036.2858535]
41. Gao L, Zhou G, Luo J, Huang Y. Word embedding with Zipf’s context. IEEE Access 2019;7:168934-168943. [doi: 10.1109/access.2019.2954691]
42. Sigurd B, Eeg-Olofsson M, van Weijer J. Word length, sentence length and frequency - Zipf revisited. Studia Linguist 2004 Apr;58(1):37-52. [doi: 10.1111/j.0039-3193.2004.00109.x]
43. Brysbaert M, Stevens M, Mandera P, Keuleers E. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. Front Psychol 2016;7:1116 [FREE Full text] [doi: 10.3389/fpsyg.2016.01116] [Medline: 27524974]
44. Goulden R, Nation P, Read J. How large can a receptive vocabulary be? Appl Linguist 1990 Dec;11(4):341-363. [doi: 10.26686/wgtn.12560441]
45. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2014; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]
46. Elfaik H, Nfaoui E. Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text. J Intell Syst 2021;30(1):395-412. [doi: 10.1515/jisys-2020-0021]
47. Suprem A, Pu C. Evaluating generalizability of fine-tuned models for fake news detection. arXiv Preprint posted online May 15, 2022 . [doi: 10.48550/arXiv.2205.07154]

48. Cui L, Lee D. CoAID: COVID-19 healthcare misinformation dataset. arXiv Preprint posted online May 22, 2020 [FREE Full text]
49. Kolluri NL, Murthy D. CoVerifi: a COVID-19 news verification system. Online Soc Netw Media 2021 Mar;22:100123 [FREE Full text] [doi: 10.1016/j.osnem.2021.100123] [Medline: 33521412]
50. Elhadad M, Li K, Gebali F. COVID-19-FAKES: a Twitter (Arabic/English) dataset for detecting misleading information on COVID-19. In: Advances in Intelligent Networking and Collaborative Systems. Manhattan, NY: Springer International; 2021 Presented at: International Conference on Intelligent Networking and Collaborative Systems; 2020; Victoria p. 256-268. [doi: 10.1007/978-3-030-57796-4_25]
51. Silva M, Ceschin F, Shrestha P, Brant C, Fernandes J, Silva CS, et al. Predicting misinformation and engagement in COVID-19 Twitter discourse in the first months of the outbreak. Association for Computing Machinery 2020 Nov;37(4):1-24 [FREE Full text]
52. Maynard D, Greenwood M. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. 2014 Presented at: Ninth International Conference on Language Resources and Evaluation (LREC'14); May 2014; Reykjavik, Iceland URL: http://www.lrec-conf.org/proceedings/lrec2014/index.html
53. Arif N, Al-Jefri M, Bizzi I, Perano G, Goldman M, Haq I, et al. Fake news or weak science? Visibility and characterization of antivaccine webpages returned by Google in different languages and countries. Front Immunol 2018 Jun 5;9:1215 [FREE Full text] [doi: 10.3389/fimmu.2018.01215]
54. Paschen J. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. J Prod Brand Manag 2019 May 06;29(2):223-233. [doi: 10.1108/jpbm-12-2018-2179]
55. Dey A, Rafi R, Hasan PS, Arko S, Chakrabarty A. Fake news pattern recognition using linguistic analysis. 2018 Presented at: Joint 7th International Conference on Informatics, Electronics Vision (ICIEV) and 2nd International Conference on Imaging, Vision Pattern Recognition (IcIVPR); June 25-29, 2018; Kitakyushu, Japan. [doi: 10.1109/iciev.2018.8641018]
56. Alonso MA, Vilares D, Gómez-Rodríguez C, Vilares J. Sentiment analysis for fake news detection. Electronics 2021 Jun 05;10(11):1348. [doi: 10.3390/electronics10111348]
57. Ajao O, Bhowmik D, Zargari S. Sentiment aware fake news detection on online social networks. 2019 Presented at: ICASSP 2019: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing; May 12-May 17, 2019; Brighton, UK p. 2507-2511. [doi: 10.1109/icassp.2019.8683170]
58. Lee T. The global rise of “fake news” and the threat to democratic elections in the USA. Public Adm Policy 2019 Jul 01;22(1):15-24. [doi: 10.1108/pap-04-2019-0008]
59. Chen E, Chang H, Rao A, Lerman K, Cowan G, Ferrara E. COVID-19 misinformation and the 2020 U.S. presidential election. HKS Misinformation Review. 2021 Mar 3. URL: https://misinforeview.hks.harvard.edu/article/covid-19-misinformation-and-the-2020-u-s-presidential-election/ [accessed 2022-09-13]
60. Agarwal I, Rana D, Panwala K, Shah R, Kathiriya V. Analysis of contextual features’ granularity for fake news detection. Research Square Preprint posted online May 27, 2022 . [doi: 10.21203/rs.3.rs-1676328/v1]
61. Sekulić I, Gjurković M, Šnajder J. Not just depressed: bipolar disorder prediction on reddit. aclanthology Preprint posted online October, 2018 . [doi: 10.18653/v1/w18-6211]
62. Deb K, De H, Chatterjee S, Pal A. Studying borderline personality disorder using machine learning. 2022 Presented at: 16th International Conference on Ubiquitous Information Management and Communication (IMCOM); January 3-5, 2022; Seoul, South Korea. [doi: 10.1109/imcom53663.2022.9721800]
63. Jubair F, Salim N, Al-Karadsheh O, Hassona Y, Saifan R, Abdel-Majeed M. Sentiment analysis for Twitter chatter during the early outbreak period of COVID-19. 2021 Presented at: 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI); December 12-17, 2021; Yogyakarta, Indonesia. [doi: 10.1109/isriti54043.2021.9702837]
64. Santos R, Pedro G, Leal S. Measuring the impact of readability features in fake news detection. 2020 Presented at: 12th Language Resources and Evaluation Conference; May 11-16, 2020; Marseille, France p. 1404-1413 URL: https://aclanthology.org/2020.lrec-1.176
65. Zhou X, Jain A, Phoha VV, Zafarani R. Fake news early detection: a theory-driven model. Digit Threats Res Pract 2020 Jun 30;1(2):1-25. [doi: 10.1145/3377478]
66. Raj C, Meel P. ARCNN framework for multimodal infodemic detection. Neural Netw 2022 Feb;146:36-68. [doi: 10.1016/j.neunet.2021.11.006] [Medline: 34839091]
67. Dementieva D, Panchenko A. Cross-lingual evidence improves monolingual fake news detection. 2021 Presented at: 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop; August 2021; Online p. 310-320. [doi: 10.18653/v1/2021.acl-srw.32]
68. Wahle J, Ashok N, Raus T, Meuschke N, Ghosal T, Gipp B. Testing the generalization of neural language models for COVID-19 misinformation detection. In: Information for a Better World: Shaping the Global Future. iConference 2022. Lecture Notes in Computer Science, Vol 13192. Cham: Springer; 2022 Presented at: Information for a Better World: Shaping the Global Future. iConference 2022; 2022; Virtual p. 381-392. [doi: 10.1007/978-3-030-96957-8_33]
69. Colas F, Brazdil P. Comparison of SVM and some older classification algorithms in text classification tasks. 2006 Presented at: Artificial Intelligence in Theory and Practice, IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream; August 21-24, 2006; Santiago, Chile p. 169-178. [doi: 10.1007/978-0-387-34747-9_18]

70. Sagi O, Rokach L. Ensemble learning: a survey. WIREs Data Mining Knowl Discov 2018 Feb 27;8(4):e1249. [doi: 10.1002/widm.1249]
71. Samek W, Wiegand T, Müller K. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv Preprint posted online August 28, 2017 [FREE Full text]
72. Roth Y, Pickles N. Updating Our Approach to Misleading Information. URL: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information [accessed 2021-07-03]
73. Meta Business Help Centre. About Fact-Checking on Facebook. URL: https://www.facebook.com/business/help/2593586717571940 [accessed 2022-09-13]
74. Nakov P, Corney D, Hasanain M. Automated fact-checking for assisting human fact-checkers. arXiv Preprint posted online March 13, 2021 [FREE Full text] [doi: 10.24963/ijcai.2021/619]
75. Bot Sentinel. More than just bots. URL: https://botsentinel.com/info/about [accessed 2022-09-13]

Abbreviations

ARCNN: Allied Recurrent and Convolutional Neural Network
ARI: automated readability index
AUC: area under the curve
BiGRU: bidirectional gated recurrent unit
BiLSTM: bidirectional long short-term memory
CNN: convolutional neural network
FN: false negative
FP: false positive
GloVE: Global Vectors for Word Representation
GRU: gated recurrent unit
KNN: K-nearest neighbor
LR: logistic regression
LSTM: long short-term memory
ML: machine learning
NB: naive Bayes
NLP: natural language processing
SVM: support vector machine
TN: true negative
TP: true positive
VADER: Valence Aware Dictionary and sEntiment Reasoner
XGBoost: extreme gradient boosting

Edited by T Mackey; submitted 19.04.22; peer-reviewed by W Ceron, SF Tsao; comments to author 09.06.22; revised version received 25.08.22; accepted 10.09.22; published 22.09.22

Please cite as:
Zhan K, Li Y, Osmani R, Wang X, Cao B
Data Exploration and Classification of News Article Reliability: Deep Learning Study
JMIR Infodemiology 2022;2(2):e38839
URL: https://infodemiology.jmir.org/2022/2/e38839
doi: 10.2196/38839
PMID: 36193330