big data and cognitive computing

MDPI

Article

Machine Learning-Based Identifications of COVID-19 Fake News Using Biomedical Information Extraction

Faizi Fifita [1], Jordan Smith [2], Melissa B. Hanzsek-Brill [2], Xiaoyin Li [2] and Mengshi Zhou [2,*]

[1] Department of Computer Science and Information Technology, St. Cloud State University, 720 4th Ave South, St. Cloud, MN 56301, USA
[2] Department of Mathematics and Statistics, St. Cloud State University, 720 4th Ave South, St. Cloud, MN 56301, USA
* Correspondence: mengshi.zhou@stcloudstate.edu

Abstract: The spread of fake news related to COVID-19 is an infodemic that leads to a public health crisis. Therefore, detecting fake news is crucial for an effective management of the COVID-19 pandemic response. Studies have shown that machine learning models can detect COVID-19 fake news based on the content of news articles. However, the use of biomedical information, which is often featured in COVID-19 news, has not been explored in the development of these models. We present a novel approach for predicting COVID-19 fake news by leveraging biomedical information extraction (BioIE) in combination with machine learning models. We analyzed 1164 COVID-19 news articles and used advanced BioIE algorithms to extract 158 novel features. These features were then used to train 15 machine learning classifiers to predict COVID-19 fake news. Among the 15 classifiers, the random forest model achieved the best performance with an area under the ROC curve (AUC) of 0.882, which is 12.36% to 31.05% higher compared to models trained on traditional features. Furthermore, incorporating BioIE-based features improved the performance of a state-of-the-art multi-modality model (AUC 0.914 vs. 0.887). Our study suggests that incorporating biomedical information into fake news detection models improves their performance, and thus could be a valuable tool in the fight against the COVID-19 infodemic.

Keywords: COVID-19; fake news; public health infodemic; machine learning; biomedical information extraction

Citation: Fifita, F.; Smith, J.; Hanzsek-Brill, M.B.; Li, X.; Zhou, M. Machine Learning-Based Identifications of COVID-19 Fake News Using Biomedical Information Extraction. *Big Data Cogn. Comput.* **2023**, *7*, 46. https://doi.org/10.3390/bdcc7010046

Academic Editor: Salvador García López

Received: 19 January 2023
Revised: 17 February 2023
Accepted: 3 March 2023
Published: 7 March 2023

1. Introduction

The emergence of the novel coronavirus (SARA-Cov-2) in December 2019 has led to the first global pandemic in which technology and social media are being used on a massive scale [1]. The impact of the coronavirus disease 2019 (COVID-19) pandemic depends on the quality of the information to which people are exposed [2]. Unfortunately, fake news about COVID-19 has traveled faster than the virus itself through various online platforms [3]. The misinformation from COVID-19 fake news can lead to negative consequences such as vaccine hesitancy [4,5], hate crimes [6,7], and psychological disorders [8–10]. The World Health Organization (WHO) has declared the spread of fake news about COVID-19 an infodemic in public health [11,12].

Containing the spread of fake news is crucial to the management of the COVID-19 pandemic response [13–15]. COVID-19 fake news can be detected by human domain experts, such as journalists or scientists [16,17]. However, detecting fake news by humans becomes resource-extensive and impossible as a massive amount of information floods the internet on a daily basis [1]. Alternatively, machine learning models can automatically evaluate the credibility of COVID-19 news from online platforms and have consequently attracted more and more attention [18,19]. Various supervised learning approaches such as the random forest [20–24], logistic regression [20,22,24], neural network [16,20,23], K nearest

neighbors [20,22], and support vector machine [20–22] were adopted to train prediction models for detecting COVID-19 fake news.

The key to machine learning-based COVID-19 fake news predictions is to extract machine-understandable features from the news articles [24,25]. Term Frequency–Inverse Document Frequency (TF–IDF) and word embedding methods are commonly used for feature extractions [16,20,22,26–31], but are often challenged by their interpretability [32,33]. Recent studies have shown that linguistic and sentiment features can be used to detect COVID-19 fake news (Figure 1). For instance, the number of uppercase characters can be used to identify the writing style of fake news [21,24], while text polarity and the count of motion words can be used to analyze the sentiment expressed in news articles [24,34,35]. However, a limitation of these features is that fake news can potentially mimic true news by parodying the writing style of real news articles and by adding fake information to authentic news pieces or by modifying the information in an authentic news article [34,36].

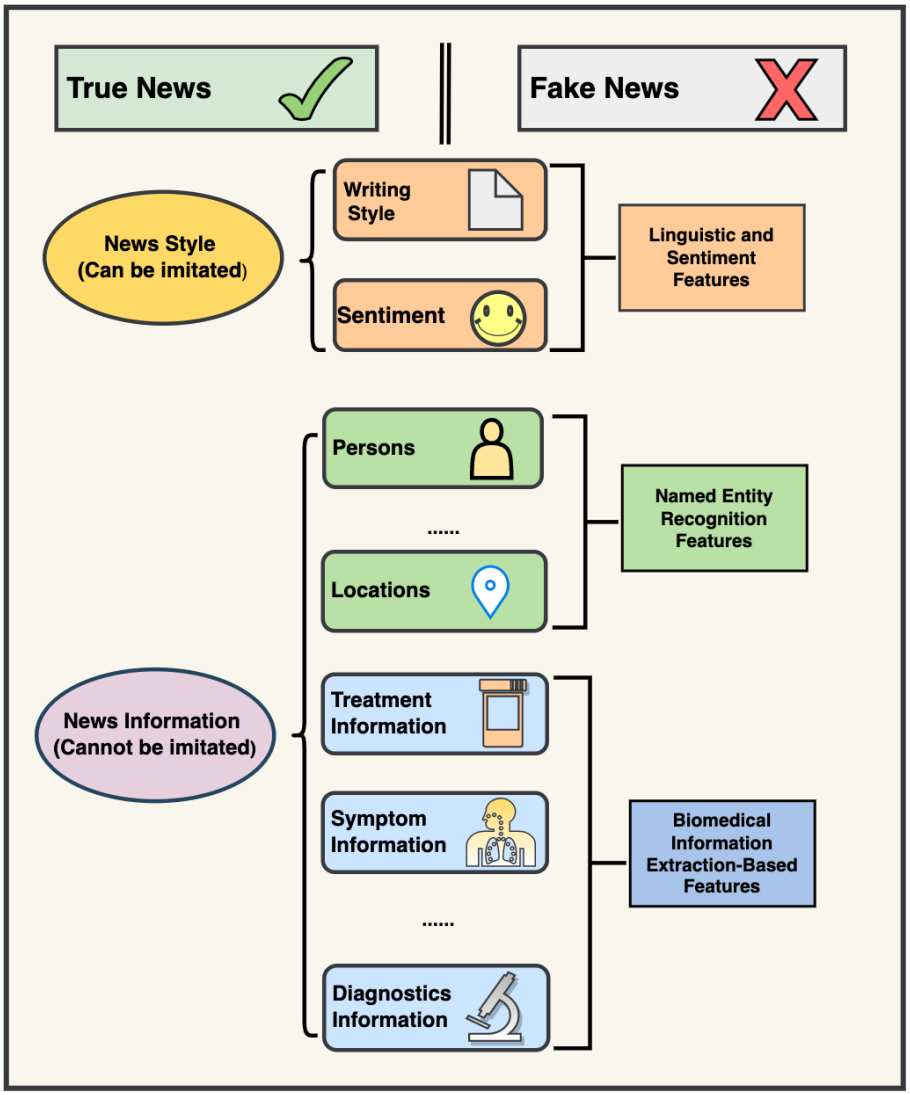

Figure 1. Machine learning features for COVID-19 news detection and the contents of news articles modeled by the features.

Information features, such as person and location names, can be used to capture specific information reported in COVID-19 news (Figure 1). Information reported in the news is less easily imitated than writing style and sentiment [37], and therefore has greater potential for identifying COVID-19 fake news. Only a limited number of studies have

incorporated information features in COVID-19 fake news prediction models [24,25,38]. Khan Suleman et al. constructed information features by extracting 18 types of predefined entities, such as the name of organizations and locations [24,25]. Similar features have been used by Gupta Ayush et al., to train machine learning algorithms [38].

Current information features have limitations in capturing biomedical information such as drug treatments, disease symptoms, and medical procedures (Figure 1). Biomedical information is a major component of COVID-19 news, with approximately half of the misinformation about COVID-19 relating to biomedical topics [39–41]. For example, former President Donald Trump and Brazilian President Jair Bolsonaro falsely claimed that hydroxychloroquine is effective as a treatment for COVID-19 [15]. Given the importance of biomedical information in COVID-19 news, incorporating such information into machine learning models can provide novel insights and potentially improve the performance of fake news prediction models. This motivates us to explore new methods or features for modeling biomedical information in existing machine learning models.

Biomedical information extraction (BioIE) aims to automatically unlock structured biomedical semantics (e.g., entities, relations, and events) out of unstructured text data [42]. BioIE has been successfully applied in drug discovery [43–45], identification of disease mechanisms [46–48], and clinical decision support [49,50]. Up to now, several tools have been developed for BioIE [51–55]. Those tools can extract thousands and hundreds of biomedical semantics, which provide us with valuable information for COVID-19 fake news detection. In order to overcome the limitation of the current information features-based methods, we proposed a new method that combines BioIE-based feature extraction and machine learning to predict COVID-19 fake news. Table 1 presents a comparison of features used to train the models in different studies aimed at detecting fake COVID-19 news. The table highlights the unique advantage of our study, which leverages the utilization of biomedical information in news articles to build COVID-19 fake news detection models.

Table 1. Comparison of features used for COVID-19 fake news detection in different studies.

Studies	Data Sources	Linguistics	Sentiment	NER	Biomedical
Alenezi et al. [26]	Twitter, WHO, CDC, etc.	No	No	No	No
Tashtoush et al. [16]	WHO, UN, Google Fact Check, etc.	No	No	No	No
Bangyal et al. [22]	Facebook, Instagram, etc.	No	No	No	No
Endo et al. [23]	Brazilian Ministry of Health, Boatos.org, etc.	Yes	No	No	No
Al-Rakhami et al. [21]	Twitter	Yes	No	No	No
Daley et al. [35]	Politifact.com	Yes	Yes	No	No
Gupta et al. [38]	Twitter	Yes	No	Yes	No
Iwendi et al. [25]	Facebook, Twitter, The New York Times, etc.	Yes	Yes	Yes	No
Khan et al. [24]	Facebook, Twitter, The New York Times, etc.	Yes	Yes	Yes	No
Our study	Facebook, Twitter, The New York Times, etc.	Yes	Yes	Yes	Yes

Linguistics: linguistics features; Sentiment: sentiment features; NER: named entity recognition features; and Biomedical: biomedical information extraction-based features.

We utilized the advanced BioIE algorithms to extract 158 BioIE-based features from over 1000 COVID-19-related news articles. The BioIE-based features were selected for the task of COVID-19 fake news detection due to their relevance and significance in capturing the biomedical information that is commonly present in news articles related to the pandemic. These features have been extracted to capture and represent the key aspects of biomedical information, including disease symptoms, treatments, and medical procedures. Our hypothesis was that incorporating these BioIE-based features into the machine learning models would enhance the prediction accuracy of fake news detection by providing additional knowledge to the models.

To verify our hypothesis, we trained machine learning classifiers to predict COVID-19 fake news using the BioIE-based features. Through rigorous evaluation processes, we provide concrete evidence that incorporating biomedical information for COVID-19 fake news prediction is a feasible and practical method with high performance. (1) BioIE-

based features are applicable when other features are not presented. (2) Models trained with BioIE-based features achieved higher performance than state-of-the-art information features-based methods. (3) A novel information features-based model, by integrating biomedical information, can achieve a higher performance than models that are trained with linguistics and semantic features. (4) A BioIE-driven multi-modality machine learning model outperformed a state-of-the-art multi-modality model. To the best of our knowledge, this study represents the first study to incorporate BioIE with machine learning-based COVID-19 fake news detections.

2. Materials and Methods

The schema for the experiment steps is shown in Figure 2. We curated a COVID-19 news dataset with known labels for true and fake news articles. We used various BioIE tools to extract biomedical information from these articles, resulting in 158 novel BioIE-based features. We then built and tested 15 supervised machine learning models based on these features. Next, we compared the performance of the BioIE-based features with state-of-the-art information features. We also combined the BioIE-based feature with existing information features to evaluate if the performance can be significantly improved. A novel information features-based model was built and compared with existing models using linguistics and semantic features. Finally, we conducted a BioIE-driven multi-modality machine learning model by integrating multiple types of features.

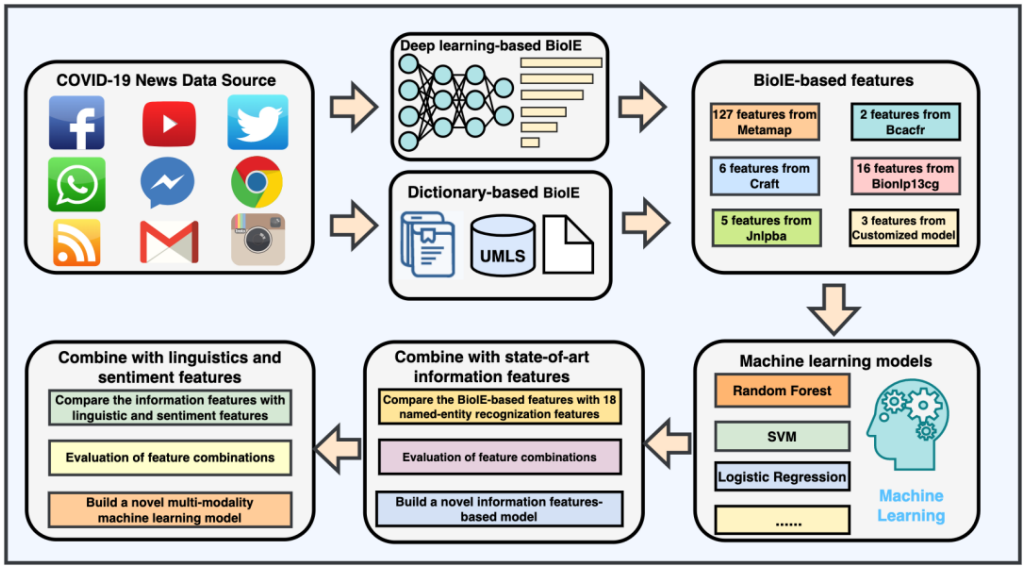

Figure 2. Flowcharts of the study.

2.1. Data

The dataset used in this work consists of 1164 COVID-19-related news articles collected from various platforms such as Twitter, Facebook, The New York Times, Harvard Health Publishing, WHO, etc. The dataset includes 586 true news articles and 578 fake news articles [56]. Previous studies have used this dataset for feature extraction and training machine learning models for detecting fake COVID-19 news [24,25]. Before building the machine learning models, we preprocessed the text of the news articles by removing URLs, punctuation marks, and empty columns. We then extracted features from the preprocessed text to build the models.

2.2. Constructing Novel Features with Biomedical Information Extraction

The 158 novel BioIE-based features were constructed through the utilization of three BioIE tools: scispaCy, MetaMap, and a spaCy customized model. These tools were applied to identify and extract various types of biomedical named entities, including diseases,

syndromes, drugs, and pathogens, from the news articles. The number of occurrences of each unique type of biomedical named entity in the news article was then appended as a feature. This process resulted in 158 unique BioIE-based features, which are listed in Tables S1–S3 in the Supplementary Materials. Each news article was recorded with its number of occurrences of each type of biomedical named entity as features for the training process.

We first performed biomedical information extraction with two existing tools, which are popular resources for recognizing biomedical named entities—scispaCy [55] and MetaMap [52]. ScispaCy is a specialized natural language processing (NLP) library that contains pre-trained deep learning models to process biomedical and clinical text [57]. MetaMap is a dictionary-based system for recognizing biomedical concepts from the Unified Medical Language System (UMLS), which is a large biomedical thesaurus that integrates nearly 200 vocabularies [58].

We used four scispaCy (version: 0.5.1) models (en_ner_craft_mb, en_ner_jnlpba_md, en_ner_bc5cdr_md, and en_ner_bionlp13cg_md) pertained on four corpuses (CRAFT [59], JNLPBA [60], BC5CDR [61], and BIONLP13CG [62]) to extract biomedical named entities. The 29 types of biomedical entities extracted from scispaCy are presented in Table S1. The number of occurrences of each type of biomedical entity represents a machine learning feature.

Unlike scispaCy, MetaMap uses a dictionary-based approach and does not require annotated data for training. We extracted 127 different types of biomedical entities in UMLS using MetaMap (V 2016). The number of occurrences of each type of biomedical entity was then recorded as a machine learning feature (Table S2).

We utilized spaCy version 3.4.0 to train a customized BioIE model on pre-tagged medical text obtained from Wikipedia [63]. The model was trained using spaCy's customized named entity recognition, which employs convolutional neural networks to develop general NLP tools for every step of the pipeline [64]. The starting point for our customized NER model was a blank spaCy NER model. We trained the model on the pre-tagged data by shuffling the training data, creating batches and training on them, and updating the model based on the training results. The model underwent 1000 iterations of training with a drop rate of 0.3 and a dynamic minibatch size, which was determined by combining the training data with an initial size of 4, a maximum size of 64, and a growth factor of 1.2. This customized model was able to extract an additional 3 features, each representing the number of occurrences of a unique type of biomedical named entity (Table S3).

2.3. Training and Evaluation with Biomedical Information Extraction-Based Feature

The BioIE-based features with the dataset in Section 2.1 were used to build classification models with machine learning algorithms. We trained 15 different machine learning configurations to predict COVID-19 fake news. To ensure the robustness of our results, we implemented regularization techniques and utilized ensemble learning algorithms to address overfitting. In addition, we randomly split the dataset into a training set (75%) and a testing set (25%) to maintain the independence and generalizability of our results. The training set was used to build and train the machine learning models to detect COVID-19 fake news, while the testing set was used to evaluate their performance. The optimization of the models' hyperparameters was performed on the training set using the successive halving method. The performance of the models was evaluated at each iteration of the hyperparameter tuning process using the area under the receiver operating characteristic curve (AUC) through a 5-fold cross-validation approach. Next, models with the best hyperparameter combinations were trained on the entire training set and applied to the testing set. The performance of these models on the testing set was measured and compared using the AUC. The accuracy, F1 score, recall, specificity, and precision were also reported. Table 2 presents the 15 machine learning configurations used in this study, along with the hyperparameters tuning range for each configuration.

Table 2. Machine learning configurations and hyperparameters tuning range.

Models	Hyperparameters	Levels
Logistic regression	Inverse of regularization strength and norm of the penalty	0.1, 1.0, and 10.0 L1 and L2
AdaBoost	Maximum number of estimators and learning rate	50, 100, and 200 0.1, 0.5, and 1.0
Bagging	Maximum number of estimators	10, 50, and 100
Decision tree	Minimum samples to split and minimum samples to be at a leaf	2.0, 5.0, and 10.0 1.0, 2.0, and 4.0
SVC_rbf	Regularization parameter	0.1, 1.0, 5.0, and 10
SVC_poly	Regularization parameter	0.1, 1.0, 5.0, and 10
SVC_sigmoid	Regularization parameter	0.1, 1.0, 5.0, and 10
kNN	Number of neighbors and weight function	3.0, 5.0, 7.0, and 9.0 Uniform and distance
Naïve bayes	Additive smoothing parameter	1.0, 5.0, and 10.0
Random forest	Number of trees and number of features	50, 100, and 200, Sqrt and Log2
SGDClassifier_L1	Learning rate	Constant, optimal, and invscaling
SGDClassifier_L2	Learning rate	Constant, optimal, and invscaling
SGDClassifier_EN	Learning rate	Constant, optimal, and invscaling
LinearSVC_L1	Regularization parameter	0.1, 1.0, 10.0, and 100.0
LinearSVC_L2	Regularization parameter	0.1, 1.0, 10.0, and 100.0

AdaBoost: adaptive boosting; Bagging: bootstrap aggregating; SVC_rbf or _poly or _sigmoid: support vector machine with rbf kernel or polynomial kernel or sigmoid kernel; SGDClassifier_L1 or _L2 or _EN: stochastic gradient descent with L1 or L2 or elastic net regularization; KNN: k-nearest neighbors; and LinearSVC_L1 or _L2: support vector machine with linear kernel coupled with L1 or L2 regularization.

The analysis was conducted using Python 3.8.8. Machine learning classifiers, random split, and evaluation metrics were conducted using Sci-kit Learn 0.24.1 [65]. We compared the prediction power of BioIE-based features extracted from different tools and evaluated the performance with the combination of those BioIE-based features. Random forest showed the best performance in our dataset. Hence, the random forest algorithm was utilized in the following analysis.

2.4. Training and Evaluation through a Combination of State-of-the-Art Information Features and Biomedical Information Extraction-Based Features

The main goal of this part is to investigate if incorporating biomedical information can improve information features-based COVID-19 fake news detection. Only a limited number of studies have extracted information features for detecting COVID-19 fake news. The 18 named entity recognition (NER) features (Table S4) adopted by Khan Suleman et al. [24,25] were used as state-of-the-art information features. Those NER features capture information in the news articles, such as locations and person names. Its limitation is that none of them preserve biomedical information. The performance of a random forest model trained with NER features is used as the baseline of information features-based COVID-19 fake news detection.

We extracted the 18 NER features with spaCy 3.4.0 Python tool kit. The NER features were validated with the same random split, hyperparameter tuning process, and evaluation method, as in Section 2.2. We compared the prediction power of state-of-the-art NER features and novel BioIE-based features. The performance of the combination of NER and

BioIE-based features was also evaluated. A novel information features-based machine learning model was developed by combining NER and BioIE-based features.

2.5. A Novel Biomedical Information Extraction-Driven Multi-Modality Machine Learning Model

Previous studies have not compared the performance of the information features and other types of features. In this part, the novel information features-based model in Section 2.3 was compared with the machine learning configurations trained with linguistics and sentiment features. The combination of linguistics and sentiment features was also compared with BioIE-based features and NER features separately. We used the state-of-the-art linguistics and sentiment features from the previous studies [24,25]. We did not include the source of the news as a feature because our study focuses on detecting COVID-19 fake news based solely on news content. We extracted 15 linguistics features (Table S5) and 5 sentiment features (Table S6) from the news articles. The same random split, hyperparameter tuning process and evaluation method in Section 2.2 were applied in this section.

The information features (BioIE-based and NER features) were then combined with linguistics and sentiment features to build a novel multi-modality machine learning model for COVID-19 fake news detection. The novel model was compared with a state-of-the-art multi-modality model recently developed by Khan Suleman et al. [24]. Khan Suleman et al. trained a multi-modality random forest model by combining linguistics, sentiments, and NER features [24]. We trained the multi-modality random forest model by further incorporating BioIE-based features. We hypothesized that integrating BioIE-based features could further improve the performance of machine learning configurations when all other types of features were presented. Both our model and the state-of-the-art model were trained on the training set and evaluated on the testing set. The same hyperparameter tuning process in Section 2.2 was applied in this section.

3. Results

3.1. Biomedical Information Extraction Is Useful for COVID-19 Fake News Detection

We first built and validated the COVID-19 fake news prediction models from biomedical information in this section. Our dataset includes a total number of 1164 COVID-19-related news. The 1164 news articles were divided into 586 true and 578 fake groups according to their label in the dataset (details in Section 2.1). First, the BioIE-based features were built using the biomedical semantics types in scispaCy, MetaMap, and the customized spaCy model, separately and combined (details in Section 2.2). Next, the list of news was randomly split into a mutually exclusive training set (75%) and a testing set (25%) with stratification from the labels. The training set was utilized for building machine learning configurations, while the testing dataset was used for evaluating the configurations' performances. The results of hyperparameter tuning are shown in Table 3.

The area under the ROC curves for the fifteen classifiers on the testing set is presented in Table 4. The accuracy, F1 score, recall, specificity, and precision of each machine learning classifier are presented in Tables S7–S11. Overall, the random forest model performed the best on all types of BioIE-based features. The random forest model on BioIE-based features extracted by all the BioIE tools achieved the highest performance (AUC = 0.882). We also observed that most of the machine configurations trained with BioIE-based features achieved an AUC higher than 0.700. The results suggested that our novel BioIE-based features have a prediction power for COVID-19 fake news detection.

Table 3. Optimal hyperparameter combinations for BioIE-based features.

Models	Hyperparameters	ScispaCy	MetaMap	Custom	Combine 1	Combine 2	Combine 3	All
Logistic regression	Inverse of regularization strength and norm of the penalty	10 L2	0.1 L2	0.1 L2	0.1 L2	1 L2	0.1 L2	0.1 L2
AdaBoost	Maximum number of estimators and learning rate	200 1.0	100 1.0	50 0.1	200 0.5	100 1.0	50 1.0	50 1.0
Bagging	Maximum number of estimators	100	50	10	100	100	100	100
Decision tree	Minimum samples to split and minimum samples to be at a leaf	10 4	10 4	2 2	10 2	10 4	5 4	10 4
SVC_rbf	Regularization parameter	5	5	5	1	1	5	1
SVC_poly	Regularization parameter	5	1	10	1	10	1	1
SVC_sigmoid	Regularization parameter	0.1	0.1	1	0.1	0.1	0.1	0.1
kNN	Number of neighbors and weight function	9 distance	9 distance	5 uniform	9 distance	9 distance	9 distance	7 distance
Naïve bayes	Additive smoothing parameter	10.0	1.0	10.0	1.0	10.0	1.0	1.0
Random forest	Number of trees and number of features	100 Log2	200 Log2	100 Sqrt	200 Log2	200 Log2	200 Sqrt	200 Sqrt
SGDClassifier_L1	Learning rate	constant	optimal	invscaling	optimal	optimal	optimal	optimal
SGDClassifier_L2	Learning rate	constant	optimal	invscaling	constant	constant	optimal	optimal
SGDClassifier_EN	Learning rate	optimal	invscaling	invscaling	optimal	invscaling	constant	optimal
LinearSVC_L1	Regularization parameter	1.0	1.0	1.0	1.0	1.0	1.0	1.0
LinearSVC_L2	Regularization parameter	1.0	1.0	1.0	1.0	1.0	1.0	1.0

ScispaCy: features extracted by scispaCy; MetaMap: features extracted by MetaMap; Custom: features extracted by the spaCy customized model; Combine 1: features extracted by scispaCy and MetaMap; Combine 2: features extracted by scispaCy and the spaCy customized model; Combine3: features extracted by MetaMap and the spaCy customized model; and All: features extracted by scispaCy, MetaMap, and the spaCy customized model.

Table 4. The area under the ROC curves on the testing set.

	ScispaCy	MetaMap	Custom	Combine 1	Combine 2	Combine 3	All
Logistic regression	0.712	0.824	0.629	0.821	0.722	0.825	0.826
AdaBoost	0.713	0.823	0.621	0.839	0.752	0.813	0.847
Bagging	0.736	0.863	0.645	0.853	0.764	0.852	0.849
Decision tree	0.658	0.722	0.646	0.703	0.637	0.730	0.747
SVC_rbf	0.741	0.844	0.669	0.836	0.740	0.844	0.844
SVC_poly	0.716	0.776	0.568	0.781	0.748	0.777	0.789
SVC_sigmoid	0.698	0.774	0.560	0.783	0.709	0.777	0.786
kNN	0.709	0.824	0.659	0.815	0.744	0.829	0.830
Naïve bayes	0.667	0.673	0.618	0.677	0.672	0.673	0.676
Random forest	0.738	0.879	0.668	0.877	0.775	0.882	0.882
SGDClassifier_L1	0.598	0.723	0.526	0.705	0.679	0.710	0.699
SGDClassifier_L2	0.598	0.723	0.526	0.745	0.643	0.710	0.699
SGDClassifier_EN	0.663	0.760	0.526	0.705	0.723	0.724	0.699
LinearSVC_L1	0.667	0.743	0.602	0.743	0.658	0.746	0.746
LinearSVC_L2	0.663	0.743	0.595	0.736	0.658	0.746	0.739

Random forest showed the best performance in our dataset. Hence, the random forest was utilized in the following analysis. To compare the performance of different types of BioIE-based features, we present the accuracy, AUC, F1 scores, recall, specificity, and precision of the random forest model in each type of BioIE-based features in Table 5. Through such a validation method, the features extracted by scispaCy achieved an accuracy of 66.7%, an AUC of 0.738, an F1 score of 0.681, a recall of 0.647, a specificity of 0.642, and a

precision of 0.693. The BioIE-based features extracted by the customized model, although only containing three features, also yielded a solid performance (accuracy 62.8%, AUC 0.668, F1 score 0.649, recall 0.647, specificity 0.617, and precision 0.639). Features extracted by MetaMap (accuracy 79.5%, AUC 0.879, F1 score 794, recall 0.760, specificity 0.762, and precision 0.832) showed a higher prediction power compared to features extracted by scispaCy or the customized model.

Table 5. Performance of random forest models trained with biomedical information extraction-based features.

	ScispaCy	MetaMap	Custom	Combine 1	Combine 2	Combine 3	All
Acc	0.667	0.795	0.628	0.778	0.698	0.781	0.799
AUC (ROC)	0.738	0.879	0.668	0.877	0.775	0.882	0.882
F1	0.681	0.794	0.649	0.778	0.695	0.784	0.801
Recall	0.647	0.76	0.647	0.747	0.660	0.760	0.780
Specificity	0.642	0.762	0.617	0.747	0.667	0.755	0.774
Precision	0.693	0.832	0.639	0.812	0.733	0.809	0.829

ScispaCy: features extracted by scispaCy; Meta: features extracted by MetaMap; Custom: features extracted by the spaCy customized model; Combine 1: features extracted by scispaCy and MetaMap; Combine 2: features extracted by scispaCy and the spaCy customized model; Combine 3: features extracted by MetaMap and the spaCy customized model; and All: features extracted by scispaCy, MetaMap, and the spaCy customized model.

MetaMap is able to identify and extract a wider variety of biomedical entities from the text compared to scispaCy and the customized model (127 vs. 29 vs. 3). We compared the median of the number of biomedical entities recognized by these three tools. Compared to scispaCy and the customized model, MetaMap was able to identify a larger number of biomedical entities, as suggested by Figure 3. Therefore, a potential reason for MetaMap features achieving the highest performance is that MetaMap was able to identify more comprehensive biomedical information from the news articles.

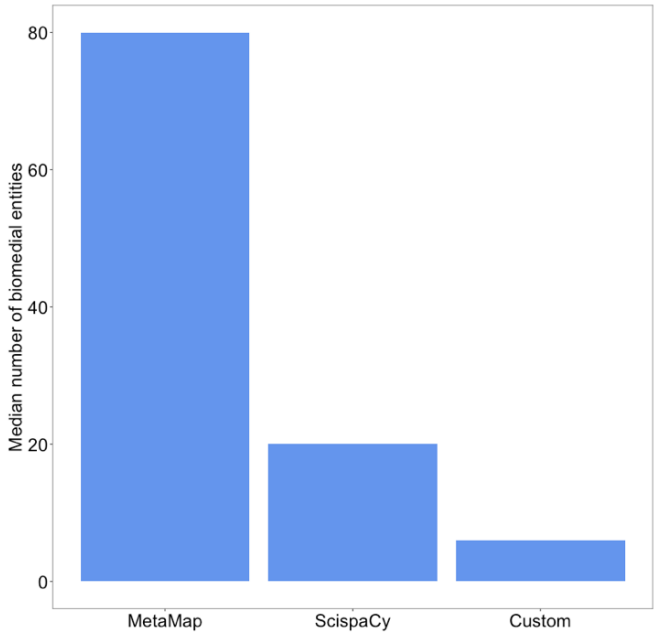

Figure 3. Median number of biomedical entities recognized by three biomedical information extraction tools.

We next tested the combination of the different types of BioIE-based features using the random forest model. The results of this section are also presented in Table 5. Incorporating features from the customized model can slightly improve the performance of scispaCy features (accuracy 69.8% vs. 66.7%, AUC 0.775 vs. 0.738, F1 score 0.695 vs. 0.681, recall 0.660

vs. 0.647, specificity 0.667 vs. 0.642, precision 0.733 vs. 0.693). The combination of features from scispaCy and MetaMap performed better than the combination of features scispaCy and the customized model (accuracy 77.8% vs. 69.8%, AUC 0.857 vs. 0.775, F1 score 0.778 vs. 0.695, recall 0.747 vs. 0.660, specificity 0.747 vs. 0.667, precision 0.812 vs. 0.733). In addition, features extracted by MetaMap can add an additional prediction power when the features extracted by scispaCy and the customized model are presented (accuracy 79.9% vs. 69.8%, AUC 0.882 vs. 0.775, F1 score 0.801 vs. 0.695, recall 0.780 vs. 0.660, specificity 0.774 vs. 0.667, precision 0.829 vs. 0.733).

Both scispaCy and the customized model extract biomedical information by deep learning-based algorithms. While MetaMap utilizes dictionary-based algorithms for BioIE. Therefore, the fact that MetaMap features can add perdition powers on scispaCy and customized features is expected. Our results showed that the BioIE-based features extracted with different models could complement each other and improve the prediction performance.

Interestingly, we found that the features extracted by scispaCy and customized models did not improve the prediction performance when the features extracted by MetaMap were presented. In machine learning model building, some features may not contribute significantly to the prediction accuracy and can even have a negative impact on the performance of the model when other features are presented. By selecting a subset of features, it is possible to remove these unimportant features and improve the performance of the model [66,67]. We hypothesized that the features from scispaCy and the customized model would not contribute significantly when combined with features from MetaMap. To verify our hypothesis, we further tested the importance of the features in the random forest model using the mean decrease in impurity [68]. Figure 4 presents the top three important features for each of the BioIE tool. Compared to features extracted by scispaCy or the customized models, the features extracted by MetaMap are more informative. This finding explains the fact that using all the features did not outperform the features extracted by MetaMap. Additionally, this suggests that the dictionary-based algorithms used by MetaMap may be more effective at extracting the relevant information from news articles.

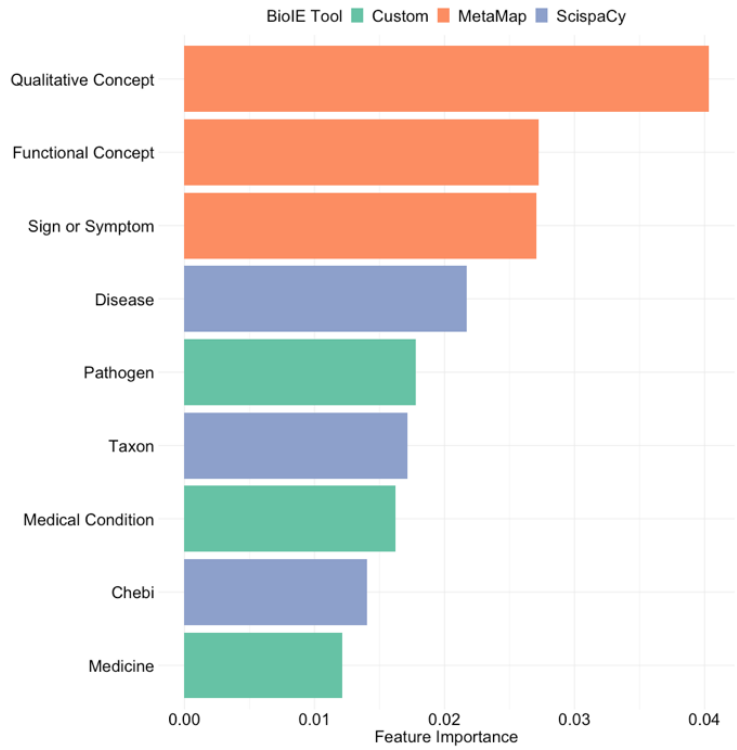

Figure 4. Top 3 important features for 3 biomedical information extraction tools.

3.2. Biomedical Information Extraction Improved the Information Features-Based COVID-19 Fake News Detection

In this section, we investigated whether BioIE-based features can improve the state-of-the-art information features-based prediction models. To set up the baseline of information features-based COVID-19 fake news detection, we trained machine learning configurations with 18 NER features employed by Khan Suleman et al. [24,25] (details in Section 2.3). We compared the performance of BioIE-based features and NER features as well as their combination in the test dataset. The random forest model was utilized in this part as it showed the best performance in Section 3.1 and in the study by Khan Suleman et al. [24,25]. The optimal hyperparameter combinations for random forest models trained on NER and BioIE-based features are presented in Table 6.

Table 6. Optimal hyperparameter combinations for random forest models trained on information features.

	Number of Trees	Number of Features
NER (baseline)	Log2	200
All	Sqrt	200
NER + ScispaCy	Sqrt	200
NER + MetaMap	Sqrt	200
NER + Custom	log2	200
NER + Combine 1	log2	200
NER + Combine 2	Sqrt	200
NER + Combine 3	log2	200
NER + All	Sqrt	200

NER: Named entity recognition features; ScispaCy: features extracted by scispaCy; MetaMap: features extracted by MetaMap; Custom: features extracted by the spaCy customized model; Combine 1: features extracted by scispaCy and MetaMap; Combine 2: features extracted by scispaCy and the spaCy customized model; Combine 3: features extracted by MetaMap and the spaCy customized model; and All: features extracted by scispaCy, MetaMap, and the spaCy customized model.

The prediction results are shown in Table 7. The accuracy, AUC, F1 score, recall, specificity, and precision in the NER features prediction in the testing set are 70.1%, 0.745, 0.709, 0.700, 0.683, and 0.719, respectively. Compared to the NER features, BioIE-based features extracted by the three BioIE tools achieved higher performances (13.98% higher in accuracy, 16.98% higher in AUC, and 12.98% higher in F1 score, 11.43% higher in recall, 13.32% higher in specificity, and 15.30% higher in precision). The results demonstrate that BioIE-based features have a stronger prediction power as compared to those state-of-the-art information features. These results also suggested that biomedical information can better describe the news information as compared to general named entities such as person and location names.

Table 7. Performance of NER and BioIE-based prediction.

	Acc.	AUC	F1	Recall	Spec.	Precision
NER (baseline)	0.701	0.754	0.709	0.700	0.683	0.719
All	0.799	0.882	0.801	0.780	0.774	0.829
NER + ScispaCy	0.781	0.830	0.765	0.727	0.732	0.807
NER + MetaMap	0.833	0.890	0.815	0.780	0.781	0.854
NER + Custom	0.743	0.807	0.753	0.753	0.732	0.753
NER + Combine 1	0.823	0.906	0.826	0.807	0.800	0.846
NER + Combine 2	0.760	0.838	0.767	0.747	0.740	0.789
NER + Combine 3	0.792	0.890	0.789	0.747	0.753	0.836
NER + All	0.819	0.899	0.814	0.787	0.784	0.843

We next built a novel information-based machine learning model by combining BioIE-based and NER features. Table 7 suggests that the combination of BioIE-based features from all three tools and NER features achieved the best performance (accuracy = 81.9%, AUC = 0.899, F1 score = 0.814, recall = 0.787, specificity = 0.784, and recall = 0.843) among

all the tested descriptors groups. This novel information features-based model improved the baseline (NER features) prediction power substantially (16.83% increase in accuracy, 19.23% increase in AUC, and 14.81% increase in F1 score, 12.42% increase in recall, 14.79% increase in specificity, and 17.25% increase in precision). Those results demonstrate that using biomedical information extraction techniques can provide more information and insights for machine learning-based COVID-19 news detection.

3.3. Information-Based Models Incorporating Biomedical Information Have a Higher Power in Identifying COVID-19 Fake News Than Linguistics-Based and Sentiment-Based Models

The predictive power of information features and other types of features have not been compared in previous studies. In this section, we investigated whether the model trained with information features can achieve a higher performance compared to those trained with linguistics and sentiment features. Table 8 presents the optimal hyperparameter combinations for random forest models trained on information, linguistics, and sentiment features. The novel information features-based model in Section 3.2 was compared with the random forest models trained with linguistics and sentiment features on the testing set (details in Section 2.3). The performances of the linguistics and sentiment features were also compared with BioIE-based features and NER features separately.

Table 8. Optimal hyperparameter combinations for random forest models trained on different types of features.

	Number of Trees	Number of Features
Linguistics	Log2	200
Sentiment	Sqrt	200
Linguistics + Sentiment	Sqrt	200
NER	Log2	200
BioIE	Sqrt	200
BioIE + NER	Sqrt	200

NER: named entity recognition features; BioIE: features extracted by scispaCy, MetaMap, and the spaCy customized model.

Table 9 compares information feature-based models with the model trained with linguistics and sentiment features. Our results show that using BioIE-based features extracted by all three BioIE tools achieves a better performance compared to using linguistics or sentiment features alone. Specifically, the AUC of BioIE-based features-based prediction was 12.36% and 31.05% higher as compared to that of linguistics or sentiment features-based predictions. Improvements in accuracy, F1 score, recall, specificity, and precision were also observed. The NER features, on the other hand, did not outperform the linguistics features. This suggests that the BioIE algorithms are more effective at extracting the relevant information from the text for the task of COVID-19 news detection.

Table 9. Comparison of information features-based models with the model trained with linguistics and sentiment features.

	Acc.	AUC	F1	Recall	Spec.	Precision
Linguistics	0.674	0.785	0.667	0.627	0.641	0.712
Sentiment	0.628	0.673	0.649	0.660	0.617	0.639
Linguistics + Sentiment	0.753	0.834	0.751	0.713	0.719	0.793
NER	0.701	0.754	0.709	0.700	0.683	0.719
BioIE	0.799	0.882	0.801	0.780	0.774	0.829
BioIE + NER	0.819	0.899	0.814	0.787	0.784	0.843

NER: named entity recognition features; BioIE: features extracted by scispaCy, MetaMap, and the spaCy customized model.

The novel information features-based model introduced in Section 3.2 (BioIE + NER in Table 9) showed a higher predictive power. Specifically, the novel information features-

based model outperformed models trained with linguistics features (accuracy 81.9% vs. 67.4%, AUC 0.899 vs. 0.785, F1 score 0.814 vs. 0.667, recall 0.787 vs. 0.627, specificity 0.784 vs. 0.641, precision 0.843 vs. 0.712), sentiment features (accuracy 81.9% vs. 62.8%, AUC 0.899 vs. 0.673, F1 score 0.814 vs. 0.649, recall 0.787 vs. 0.660, specificity 0.784 vs. 0.617, precision 0.843 vs. 0.639), and the combination of linguistics and sentiment features (accuracy 81.9% vs. 75.3%, AUC 0.899 vs. 0.0.834, F1 score 0.814 vs. 0.751, recall 0.787 vs. 0.713, specificity 0.784 vs. 0.719, precision 0.843 vs. 0.793).

The results in this section indicate that information-based models outperform the linguistics-based and sentiment-based models when incorporating biomedical information. The information reported in the news is less imitable compared to the news' linguistics style and sentiment [34,36,37,69]. Therefore, models trained with information features have an increased ability to identify COVID-19 fake news. Our study shows that addressing the limitation of current information features to preserve biomedical information allows information-based models to achieve a better performance in detecting fake COVID-19 news than linguistics-based and sentiment-based models. These findings highlight the importance of preserving biomedical information for effective fake news detection in the context of COVID-19.

3.4. A Novel Biomedical Information-Driven Multi-Modality Model Outperforms a State-of-the-Art Multi-Modality Model

The goal of this section was to build a novel multi-modality prediction model for COVID-19 fake news detection by combining BioIE-based features with the existing features (linguistics, sentiment, and NER features; details in Section 2.3). After comparing the different combinations of BioIE-based features, we used the BioIE-based features extracted by MetaMap in this section.

We compared the novel multi-modality model with a state-of-the-art multi-modality model developed by Khan Suleman et al. [24]. In their study, Khan Suleman et al., developed a multi-modality model by training random forest using the combination of linguistics, sentiment, and NER features [24]. Table 10 presents the optimal hyperparameter combinations for the novel biomedical information-driven multi-modality model and the state-of-the-art multi-modality model. Figure 5 illustrates the ROC curves of our model and the state-of-the-art model for classifying COVID-19 fake news vs. true news in the testing data. Our model achieved a higher AUC of 0.914 compared with the 0.887 obtained by the state-of-the-art model.

Table 10. Optimal hyperparameter combinations for the novel biomedical information-driven multi-modality model and the state-of-the-art multi-modality model.

	Number of Trees	**Number of Features**
BioIE-driven model	Sqrt	200
State-of-the-art model	Sqrt	200

Compared to the state-of-the-art model, our new model has the advantage of modeling biomedical information from news articles. The results indicate that machine learning models have learned complementary features for detecting fake COVID-19 news based on the BioIE-based features and existing features. This suggests that biomedical information can provide additional knowledge for machine learning-based COVID-19 detection models beyond the linguistic style, sentiment, and non-biomedical named entities.

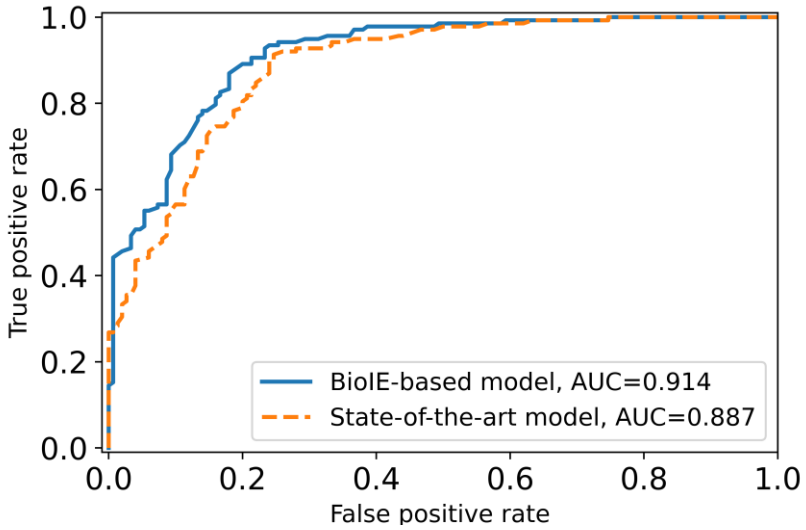

Figure 5. ROC curves for the validation test results for the BioIE-based multi-modality model and the state-of-the-art multi-modality model.

4. Discussion

The need for computational models to aid human fact-checkers is crucial, not only for the current COVID-19 pandemic but also for future unexpected infodemics. In this study, we developed a novel approach to identify COVID-19 fake news from biomedical information. We extended the scope of usage of biomedical information extraction (BioIE) to machine learning-based COVID-19 fake news detection. By using existing BioIE tools, we constructed novel machine learning features and provided a comprehensive evaluation of their performance in comparison to existing features in the context of machine learning-based COVID-19 fake news detection. Our results showed that the use of BioIE-based features alone was effective in detecting fake news. Additionally, we found that the combination of BioIE-based features and existing features further improved the performance of the COVID-19 fake news detection models. Our results demonstrate that using biomedical information extraction techniques to mine additional information from COVID-19 news articles can provide additional knowledge and insights for COVID-19 fake news prediction models. These findings suggest that biomedical information can be a valuable source for detecting COVID-19 fake news. Our study provides a comprehensive evaluation of the impact of incorporating biomedical information into machine learning models and lays the foundation for the future computational COVID-19 fake news detection models.

Our work has several limitations that could be improved in future studies. First, our machine learning-based prediction model can be improved by incorporating additional biomedical information. Currently, we only extracted biomedical entities from the news articles. Biomedical relationships, such as drug–disease treatment associations [70], disease phenotypes [71,72], and drug–side effect relationships [44,73], may provide more accurate information for fake news prediction models. For example, a biomedical relationship with "Hydroxychloroquine treats COVID-19" is more informative than biomedical entities "Hydroxychloroquine" and "COVID-19" on their own. Currently, named entity recognition (NER) features are the only existing information features used for COVID-19 fake news prediction models. Thus, we used biomedical entities to represent BioIE-based features in order to fairly compare them with NER features. Extracting biomedical relationships for building machine learning features is one of our next steps.

Second, biomedical information extraction itself remains a challenging task. Currently, scispaCy and MetaMap are the most commonly used BioIE tools. Studies suggest that the precision of scispaCy and MetaMap in biomedical named entity recognition is around 70% and 85% [52,55]. Therefore, the biomedical information extracted from the new articles may not be fully accurate. In this study, our goal is to demonstrate the potential contributions of

biomedical information extraction for COVID-19 fake news detection, rather than to build a perfect prediction model. In the future, we plan to develop new methods to enhance the accuracy of biomedical information extraction from news articles. We believe that improved BioIE models could enhance the performance of our prediction method.

In addition, the news dataset used to train the model only consists of text. Fake news can also be spread through other types of content, such as audio clips, images, and videos [74]. Currently, there are no available COVID-19 fake news datasets that include these types of content. A feature goal of this study is to generate new datasets that include a wider range of news content types, and to develop computational models that can identify COVID-19 fake news from these datasets. This will enable us to more effectively identify fake news that is spread through non-textual content.

Last but not least, our novel approach was developed based on the news collected in the English language. Therefore, the model is currently only applied to English text. However, the spread of COVID-19 fake news is a global issue, so it is important to develop machine learning models that can work in multiple languages. Recently, computational approaches have been developed that can deal with fake news in multiple languages [23,75–77], but multilingual biomedical information extraction remains a challenge. MetaMap and scispaCy can be applied only to English text [52,78]. We anticipate that our model can be expanded to non-English news with effectively multilingual biomedical information extraction methods.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bdcc7010046/s1; Table S1: Biomedical information extraction-based features extracted by scispaCy; Table S2: Biomedical information extraction-based features extracted by MetaMap; Table S3: Biomedical information extraction-based features extracted by spaCy customized model; Table S4: Named-entity recognition features; Table S5: Linguistic features; Table S6: Sentiment features; Table S7: Accuracies of the 15 machine learning configurations on the validation set; Table S8: F1 scores of the 15 machine learning configurations on the validation set; Table S9: Recalls of the 15 machine learning configurations on the testing set; Table S10: Specificities of the 15 machine learning configurations on the testing set; and Table S11: Precisions of the 15 machine learning configurations on the testing set.

Author Contributions: Study conception and supervision: M.Z. study design: M.Z.; data preprocessing: F.F.; analysis and evaluation: F.F. and J.S.; paper writing: M.Z., F.F. and J.S.; paper review and supervision: M.Z., X.L. and M.B.H.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation under grant no. 1742517 and the St. Cloud State University early career grant.

Data Availability Statement: The source code and datasets used in this study are available at https://github.com/FayZ676/COVID-nlp-research, (accessed on 18 January 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bang, Y.; Ishii, E.; Cahyawijaya, S.; Ji, Z.; Fung, P. Model generalization on COVID-19 fake news detection. In Proceedings of the International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Online, 8 February 2021; pp. 128–140.
2. Pennycook, G.; McPhetres, J.; Zhang, Y.; Lu, J.G.; Rand, D.G. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychol. Sci.* **2020**, *31*, 770–780. [CrossRef] [PubMed]
3. Radwan, E.; Radwan, A.; Radwan, W. The role of social media in spreading panic among primary and secondary school students during the COVID-19 pandemic: An online questionnaire study from the Gaza Strip, Palestine. *Heliyon* **2020**, *6*, e05807. [CrossRef] [PubMed]
4. Freeman, D.; Waite, F.; Rosebrock, L.; Petit, A.; Causier, C.; East, A.; Jenner, L.; Teale, A.-L.; Carr, L.; Mulhall, S. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychol. Med.* **2022**, *52*, 251–263. [CrossRef] [PubMed]

5. Pierri, F.; Perry, B.; DeVerna, M.R.; Yang, K.-C.; Flammini, A.; Menczer, F.; Bryden, J. The impact of online misinformation on US COVID-19 vaccinations. *arXiv* **2021**, arXiv:2104.10635.
6. Orellana, C.I. Health workers as hate crimes targets during COVID-19 outbreak in the Americas. *Rev. Salud Pública* **2020**, *22*, 253–257.
7. Kim, J.Y.; Kesari, A. Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech During the COVID-19 Pandemic. *J. Online Trust Saf.* **2021**, *1*, 1–14. [CrossRef]
8. Rocha, Y.M.; de Moura, G.A.; Desidério, G.A.; de Oliveira, C.H.; Lourenço, F.D.; de Figueiredo Nicolete, L.D. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *J. Public Health* **2021**, 1–10. [CrossRef]
9. Ahmad, A.R.; Murad, H.R. The impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan: Online questionnaire study. *J. Med. Internet Res.* **2020**, *22*, e19556. [CrossRef]
10. Secosan, I.; Virga, D.; Crainiceanu, Z.P.; Bratu, L.M.; Bratu, T. Infodemia: Another enemy for romanian frontline healthcare workers to fight during the COVID-19 outbreak. *Medicina* **2020**, *56*, 679. [CrossRef]
11. World Health Organization. Novel Coronavirus (2019-nCoV) Situation Report-13. Available online: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf (accessed on 5 December 2022).
12. Zarocostas, J. How to fight an infodemic. *Lancet* **2020**, *395*, 676. [CrossRef]
13. Bavel, J.J.V.; Baicker, K.; Boggio, P.S.; Capraro, V.; Cichocka, A.; Cikara, M.; Crockett, M.J.; Crum, A.J.; Douglas, K.M.; Druckman, J.N. Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **2020**, *4*, 460–471. [CrossRef] [PubMed]
14. Habersaat, K.B.; Betsch, C.; Danchin, M.; Sunstein, C.R.; Böhm, R.; Falk, A.; Brewer, N.T.; Omer, S.B.; Scherzer, M.; Sah, S. Ten considerations for effectively managing the COVID-19 transition. *Nat. Hum. Behav.* **2020**, *4*, 677–687. [CrossRef] [PubMed]
15. van Der Linden, S.; Roozenbeek, J.; Compton, J. Inoculating against fake news about COVID-19. *Front. Psychol.* **2020**, *11*, 566790. [CrossRef] [PubMed]
16. Tashtoush, Y.; Alrababah, B.; Darwish, O.; Maabreh, M.; Alsaedi, N. A Deep Learning Framework for Detection of COVID-19 Fake News on Social Media Platforms. *Data* **2022**, *7*, 65. [CrossRef]
17. Zhang, X.; Ghorbani, A.A. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* **2020**, *57*, 102025. [CrossRef]
18. Varma, R.; Verma, Y.; Vijayvargiya, P.; Churi, P.P. A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-COVID-19 pandemic. *Int. J. Intell. Comput. Cybern.* **2021**, *14*, 617–646. [CrossRef]
19. Choraś, M.; Demestichas, K.; Giełczyk, A.; Herrero, Á.; Ksieniewicz, P.; Remoundou, K.; Urda, D.; Woźniak, M. Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Appl. Soft Comput.* **2021**, *101*, 107050. [CrossRef]
20. Abdelminaam, D.S.; Ismail, F.H.; Taha, M.; Taha, A.; Houssein, E.H.; Nabil, A. Coaid-deep: An optimized intelligent framework for automated detecting covid-19 misleading information on twitter. *IEEE Access* **2021**, *9*, 27840–27867. [CrossRef] [PubMed]
21. Al-Rakhami, M.S.; Al-Amri, A.M. Lies kill, facts save: Detecting COVID-19 misinformation in twitter. *IEEE Access* **2020**, *8*, 155961–155970. [CrossRef]
22. Bangyal, W.H.; Qasim, R.; Ahmad, Z.; Dar, H.; Rukhsar, L.; Aman, Z.; Ahmad, J. Detection of fake news text classification on COVID-19 using deep learning approaches. *Comput. Math. Methods Med.* **2021**, *2021*, 5514220. [CrossRef]
23. Endo, P.T.; Santos, G.L.; de Lima Xavier, M.E.; Nascimento Campos, G.R.; de Lima, L.C.; Silva, I.; Egli, A.; Lynn, T. Illusion of Truth: Analysing and Classifying COVID-19 Fake News in Brazilian Portuguese Language. *Big Data Cogn. Comput.* **2022**, *6*, 36. [CrossRef]
24. Khan, S.; Hakak, S.; Deepa, N.; Prabadevi, B.; Dev, K.; Trelova, S. Detecting COVID-19-Related Fake News Using Feature Extraction. *Front. Public Health* **2021**, *9*, 788074. [CrossRef] [PubMed]
25. Iwendi, C.; Mohan, S.; Khan, S.; Ibeke, E.; Ahmadian, A.; Ciano, T. Covid-19 fake news sentiment analysis. *Comput. Electr. Eng.* **2022**, *101*, 107967. [CrossRef] [PubMed]
26. Alenezi, M.N.; Alqenaei, Z.M. Machine learning in detecting COVID-19 misinformation on twitter. *Future Internet* **2021**, *13*, 244. [CrossRef]
27. Fauzi, A.; Setiawan, E.; Baizal, Z. Hoax news detection on Twitter using term frequency inverse document frequency and support vector machine method. *J. Phys. Conf. Ser.* **2019**, *1192*, 012025. [CrossRef]
28. Kong, S.H.; Tan, L.M.; Gan, K.H.; Samsudin, N.H. Fake news detection using deep learning. In Proceedings of the 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, 18–19 April 2020; pp. 102–107.
29. Baarir, N.F.; Djeffal, A. Fake news detection using machine learning. In Proceedings of the 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being (IHSH), Boumerdes, Algeria, 9–10 February 2021; pp. 125–130.
30. Goldani, M.H.; Momtazi, S.; Safabakhsh, R. Detecting fake news with capsule neural networks. *Appl. Soft Comput.* **2021**, *101*, 106991. [CrossRef]
31. Bogale Gereme, F.; Zhu, W. Fighting fake news using deep learning: Pre-trained word embeddings and the embedding layer investigated. In Proceedings of the 2020 The 3rd International Conference on Computational Intelligence and Intelligent Systems, Tokyo, Japan, 13–15 November 2020; pp. 24–29.

32. Qaiser, S.; Ali, R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **2018**, *181*, 25–29. [CrossRef]
33. Khattak, F.K.; Jeblee, S.; Pou-Prom, C.; Abdalla, M.; Meaney, C.; Rudzicz, F. A survey of word embeddings for clinical text. *J. Biomed. Inform.* **2019**, *100*, 100057. [CrossRef] [PubMed]
34. Alonso, M.A.; Vilares, D.; Gómez-Rodríguez, C.; Vilares, J. Sentiment analysis for fake news detection. *Electronics* **2021**, *10*, 1348. [CrossRef]
35. Daley, B.P. Leveraging Machine Learning for Automatically Classifying Fake News in the COVID-19 Outbreak. 2020. Available online: https://scholarworks.boisestate.edu/icur/2020/Poster_Session/118/ (accessed on 22 August 2022).
36. Zhou, Z.; Guan, H.; Bhat, M.M.; Hsu, J. Fake news detection via NLP is vulnerable to adversarial attacks. *arXiv* **2019**, arXiv:1901.09657.
37. Lazer, D.M.; Baum, M.A.; Benkler, Y.; Berinsky, A.J.; Greenhill, K.M.; Menczer, F.; Metzger, M.J.; Nyhan, B.; Pennycook, G.; Rothschild, D. The science of fake news. *Science* **2018**, *359*, 1094–1096. [CrossRef] [PubMed]
38. Gupta, A.; Sukumaran, R.; John, K.; Teki, S. Hostility detection and covid-19 fake news detection in social media. *arXiv* **2021**, arXiv:2101.05953.
39. Brennen, J.S.; Simon, F.M.; Howard, P.N.; Nielsen, R.K. *Types, Sources, and Claims of COVID-19 Misinformation*; University of Oxford: Oxford, UK, 2020.
40. Posetti, J.; Bontcheva, K. Disinfodemic: Deciphering COVID-19 Disinformation. Policy Brief. 2020, Volume 1. Available online: https://en.unesco.org/covid19/disinfodemic/brief1 (accessed on 25 August 2022).
41. Charquero-Ballester, M.; Walter, J.G.; Nissen, I.A.; Bechmann, A. Different types of COVID-19 misinformation have different emotional valence on Twitter. *Big Data Soc.* **2021**, *8*, 20539517211041279. [CrossRef]
42. Liu, F.; Chen, J.; Jagannatha, A.; Yu, H. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv* **2016**, arXiv:1606.07993.
43. Zhou, M.; Wang, Q.; Zheng, C.; John Rush, A.; Volkow, N.D.; Xu, R. Drug repurposing for opioid use disorders: Integration of computational prediction, clinical corroboration, and mechanism of action analyses. *Mol. Psychiatry* **2021**, *26*, 5286–5296. [CrossRef] [PubMed]
44. Zhou, M.; Chen, Y.; Xu, R. A drug-side effect context-sensitive network approach for drug target prediction. *Bioinformatics* **2019**, *35*, 2100–2107. [CrossRef] [PubMed]
45. Zhou, M.; Zheng, C.; Xu, R. Combining phenome-driven drug-target interaction prediction with patients' electronic health records-based clinical corroboration toward drug discovery. *Bioinformatics* **2020**, *36*, i436–i444. [CrossRef] [PubMed]
46. Pan, Y.; Xu, R. Mining comorbidities of opioid use disorder from FDA adverse event reporting system and patient electronic health records. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 1–11. [CrossRef] [PubMed]
47. Zheng, C.; Xu, R. The Alzheimer's comorbidity phenome: Mining from a large patient database and phenome-driven genetics prediction. *JAMIA Open* **2019**, *2*, 131–138. [CrossRef]
48. Zheng, C.; Xu, R. Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data. *BMC Bioinform.* **2018**, *19*, 85–93. [CrossRef]
49. Friedman, C.; Hripcsak, G.; Shagina, L.; Liu, H. Representing information in patient reports using natural language processing and the extensible markup language. *J. Am. Med. Inform. Assoc.* **1999**, *6*, 76–87. [CrossRef] [PubMed]
50. Cao, Y.; Liu, F.; Simpson, P.; Antieau, L.; Bennett, A.; Cimino, J.J.; Ely, J.; Yu, H. AskHERMES: An online question answering system for complex clinical questions. *J. Biomed. Inform.* **2011**, *44*, 277–288. [CrossRef] [PubMed]
51. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In Proceedings of the AMIA Symposium, Washington, DC, USA, 3–7 November 2001; p. 17.
52. Aronson, A.R.; Lang, F.M. An overview of MetaMap: Historical perspective and recent advances. *J. Am. Med. Inf. Assoc.* **2010**, *17*, 229–236. [CrossRef] [PubMed]
53. Tang, B.; Feng, Y.; Wang, X.; Wu, Y.; Zhang, Y.; Jiang, M.; Wang, J.; Xu, H. A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature. *J. Cheminform.* **2015**, *7*, 1–6. [CrossRef] [PubMed]
54. Leaman, R.; Wei, C.-H.; Lu, Z. tmChem: A high performance approach for chemical named entity recognition and normalization. *J. Cheminform.* **2015**, *7*, 1–10. [CrossRef] [PubMed]
55. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. ScispaCy: Fast and robust models for biomedical natural language processing. *arXiv* **2019**, arXiv:1902.07669.
56. Github. Fake News Dataset. Available online: https://raw.githubusercontent.com/susanli2016/NLP-with-Python/master/data/corona_fake.csv (accessed on 14 May 2022).
57. Hussain, S.-A.; Sezgin, E.; Krivchenia, K.; Luna, J.; Rust, S.; Huang, Y. A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: A cystic fibrosis case study. *JAMA Open* **2021**, *4*, ooab084. [CrossRef]
58. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef]
59. Bada, M.; Eckert, M.; Evans, D.; Garcia, K.; Shipley, K.; Sitnikov, D.; Baumgartner, W.A.; Cohen, K.B.; Verspoor, K.; Blake, J.A. Concept annotation in the CRAFT corpus. *BMC Bioinform.* **2012**, *13*, 1–20. [CrossRef]

60. Huang, M.-S.; Lai, P.-T.; Tsai, R.T.-H.; Hsu, W.-L. Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task. *arXiv* **2019**, arXiv:1901.10219.
61. Li, J.; Sun, Y.; Johnson, R.J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Wiegers, T.C.; Lu, Z. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database* **2016**, *2016*, baw068. [CrossRef]
62. Pyysalo, S.; Ohta, T.; Rak, R.; Rowley, A.; Chun, H.-W.; Jung, S.-J.; Choi, S.-P.; Tsujii, J.i.; Ananiadou, S. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC Bioinform.* **2015**, *16*, 1–19. [CrossRef] [PubMed]
63. Kaggle. Available online: https://www.kaggle.com/datasets/finalepoch/medical-ner (accessed on 1 June 2022).
64. Honnibal, M.; Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Appear* **2017**, *7*, 411–420.
65. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
66. Liu, H.; Motoda, H.; Setiono, R.; Zhao, Z. Feature selection: An ever evolving frontier in data mining. In Proceedings of the Feature Selection in Data Mining, Hyderabad, India, 21 June 2010; pp. 4–13.
67. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
68. Louppe, G.; Wehenkel, L.; Sutera, A.; Geurts, P. Understanding variable importances in forests of randomized trees. *Adv. Neural Inf. Process. Syst.* **2013**, *26*. Available online: https://proceedings.neurips.cc/paper/2013/hash/e3796ae838835da0b6f6ea37bcf8bcb7-Abstract.html (accessed on 6 October 2022).
69. Molina, M.D.; Sundar, S.S.; Le, T.; Lee, D. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *Am. Behav. Sci.* **2021**, *65*, 180–212. [CrossRef]
70. Xu, R.; Wang, Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC Bioinform.* **2013**, *14*, 1–11. [CrossRef]
71. Xu, R.; Li, L.; Wang, Q. Towards building a disease-phenotype knowledge base: Extracting disease-manifestation relationship from literature. *Bioinformatics* **2013**, *29*, 2186–2194. [CrossRef]
72. Xu, R.; Li, L.; Wang, Q. dRiskKB: A large-scale disease-disease risk relationship knowledge base constructed from biomedical text. *BMC Bioinform.* **2014**, *15*, 1–13. [CrossRef]
73. Xu, R.; Wang, Q. A Knowledge-Driven Approach in Constructing a Large-Scale Drug-Side Effect Relationship Knowledge Base for Computational Drug Discovery. In Proceedings of the Bioinformatics Research and Applications: 10th International Symposium, ISBRA 2014, Zhangjiajie, China, 28–30 June 2014; p. 391.
74. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **2019**, *9*, 39–52. [CrossRef]
75. Abonizio, H.Q.; de Morais, J.I.; Tavares, G.M.; Barbon Junior, S. Language-independent fake news detection: English, Portuguese, and Spanish mutual features. *Future Internet* **2020**, *12*, 87. [CrossRef]
76. Guibon, G.; Ermakova, L.; Seffih, H.; Firsov, A.; Le Noé-Bienvenu, G. Multilingual fake news detection with satire. In Proceedings of the CICLing: International Conference on Computational Linguistics and Intelligent Text Processing, La Rochelle, France, 7–13 April 2019.
77. Lee, J.-W.; Kim, J.-H. Fake Sentence Detection Based on Transfer Learning: Applying to Korean COVID-19 Fake News. *Appl. Sci.* **2022**, *12*, 6402. [CrossRef]
78. Digan, W.; Névéol, A.; Neuraz, A.; Wack, M.; Baudoin, D.; Burgun, A.; Rance, B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 504–515. [CrossRef] [PubMed]