applied sciences

MDPI

Article

Fighting the COVID-19 Infodemic in News Articles and False Publications: The NeoNet Text Classifier, a Supervised Machine Learning Algorithm

Mohammad A. R. Abdeen [1], Ahmed Abdeen Hamed [2,*] and Xindong Wu [3]

[1] Faculty of Computer and Information Systems, Islamic University of Madinah, Medina 42351, Saudi Arabia; mabdeen@iu.edu.sa
[2] School of Cybersecurity, Data Science and Computing, Norwich University, Northfield, VT 05663, USA
[3] Mininglamp Academy of Sciences, Mininglamp Technology, Beijing 100864, China; wuxindong@mininglamp.com
* Correspondence: ahamed@norwich.edu; Tel.: +1-8123602703

Abstract: The spread of the Coronavirus pandemic has been accompanied by an infodemic. The false information that is embedded in the infodemic affects people's ability to have access to safety information and follow proper procedures to mitigate the risks. This research aims to target the falsehood part of the infodemic, which prominently proliferates in news articles and false medical publications. Here, we present NeoNet, a novel supervised machine learning algorithm that analyzes the content of a document (news article, a medical publication) and assigns a label to it. The algorithm was trained by Term Frequency Inverse Document Frequency (TF-IDF) bigram features, which contribute a network training model. The algorithm was tested on two different real-world datasets from the CBC news network and COVID-19 publications. In five different fold comparisons, the algorithm predicted a label of an article with a precision of 97–99%. When compared with prominent algorithms such as Neural Networks, SVM, and Random Forests NeoNet surpassed them. The analysis highlighted the promise of NeoNet in detecting disputed online contents, which may contribute negatively to the COVID-19 pandemic.

Keywords: COVID-19 infodemic; text classification; TF-IDF features; network training modes; supervised learning; misinformation; news classification; false publications; PubMed anomaly detection

Citation: Abdeen, M.A.R.; Hamed, A.A.; Wu, X. Fighting the COVID-19 Infodemic in News Articles and False Publications: The NeoNet Text Classifier, a Supervised Machine Learning Algorithm. *Appl. Sci.* **2021**, *11*, 7265. https://doi.org/10.3390/app11167265

Academic Editor: Giancarlo Mauri

Received: 17 June 2021
Accepted: 30 July 2021
Published: 6 August 2021

1. Introduction

Without doubt, the Coronavirus pandemic has affected the world around us in unprecedented ways. Particularly, an emerging infodemic of news articles, social media posts, and publications has accompanied the global pandemic and circulated a vast volume of information, some of which is misleading [1–5]. According to the World Health Organization, an infodemic is "an overabundance of information—some accurate and some not." [6]. This means that our digital world is riddled with an enormous amount of misinformation and disinformation resulting from fake news articles, careless social media posts, or publications that have not gone through a rigorous peer-review process [7]. As a result, rumors, conspiracy theories, and stigma are linked to the ongoing COVID-19 pandemic and circulated on social media platforms and news networks. The impact of the infodemic on the general-public is unquestionable, as it makes it hard for people to identify reliable guidelines from trustworthy sources [8]. Clearly, the spread of misinformation and disinformation has existed long before the pandemic. It has also been considered as a social-determinant of health due to its impact [9].

The coronavirus infodemic aspects are numerous: (1) The spread of rumors across the world has led to inappropriate behavior and has caused an adverse effect on people's physical and the mental health [2,10]; (2) conspiracy theories have spread widely during

the pandemic in attempt to explain the unusual circumstances [11]. In fact, similar theories emerged during the SARS outbreak in China and the Ebola outbreak the Congo [12]; (3) Misinformation and public health damage has been related to tweeting bad advice from people of authority. For instance, Orso et al., stated that in a tweet, the French minister of health warned the citizens of his country not to use certain drugs (e.g., cortisone), advice that has gone viral during the pandemic [13]. Later on, clinical trials proved that cortisone is beneficial. Clearly, such events have the effect of dispensing significant treatment, and in this case, any reference to cortisone was eliminated; (4) Stigma which overwhelmed social media in the form of hashtags contributed to a backlash against countries and people (e.g., stigma against China and Chinese people) [12], (5) disinformation, which is an intentional act to deliver false information to mislead the general public. A significant instance that took place during the pandemic was the promotion of vitamin D by an Indonesian author. The article and its recommendations turned out to be from a suspicions source, as the authors' names were never linked to the listed affiliation. Such an article was downloaded 17,000 times and mentioned 8000 times on social media platforms. The matter made worse when the article was also broadcasted by DailyMail, a major news network, in an article entitled: "Terrifying chart shows how Covid-19 patients who end up in hospital may be almost certain to die if they have a vitamin D deficiency" [14]. Indeed, it is terrifying to witness major news organizations making make life and death assertions based on suspicious sources such as this suspicious kind of publication.

Presenting the above evidence begged the question of "Who do you trust? How to better mitigate?". Several efforts have also emerged to address the flood of information and provided guidelines and recommendations on how to answer such questions. In fact, this exact question has been answered an article titled: "Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic" [15]. The authors of the referenced article referred to how the digital disruption of social media and search engines are responsible for the digital destruction by the act of propagating misinformation. The article urged for the development of new methods and approaches to establish and build trust among the users and their platforms. Another prominent reference was titled: "How to fight an Infodemic: The Four Pillars of Infodemic Management" [16]. The pillars included (1) monitoring of information, (2) knowledge refinement and quality improvement processes (e.g., fact-checking), (3) the presentation of timely and accurate knowledge that minimizes or eliminates the influence of commercial and political influence, and (4) advocating for facts and science, which often have been overtaken by social media advertisements presenting "inappropriate content". Another effort that addresses the trustworthiness of online knowledge and information sources introduced a COVID-19 infodemic crowdsourcing framework [6]. The effort resulted in recommendations that also overlapped with the four pillars presented in [16] (e.g., knowledge refinement for fact-checking). The recommendations stated the importance of using computational methods such as artificial intelligence (AI) and machine learning (ML) to produce insights that enable decision making to manage the infodemic.

From the wide spectrum of issues associated with the COVID-19 infodemic and assessments [17], and recommendations made by the scholars in the field, we believe that computational scientists have a significant role to play in the fight against this infodemic. Particularly, the use of artificial intelligence, natural language processing, and machine learning must demonstrate their full potential in this fight. As demonstrated by the DailyMail news article, disinformation thrives in major news networks. The impact of such articles is clearly magnified when it is also socialized on social media platforms such as (Twitter and Facebook). It is imperative to address misinformation and disinformation at the source (i.e., the news article) before it is socialized on social media and becomes viral. A step that is dire at this point is the development of a mechanism that analyzes the content of an article to assess how viable the content of a news article is from a linguistic point of view. We argue that each news article must pass a step of label-prediction; otherwise, it must be flagged as potentially untrustworthy. This has to be accomplished by measuring the

quality of the linguistic aspects of the article. Noun-phrases, for example, are essential for making up the main facts of each article. Therefore, any computational mechanism must utilize the noun-phrases to decide if an article should pass the label-prediction process. The final outcome generates a COVID-19 safe/disputed label accordingly. In the past few months, Twitter started flagging socialized contents of political dispute "this claim is disputed". Twitter has also taken more advanced measures and applied filters to remove vaccine misinformation from the platform [18]. In the ideal scenario, we envision that our mechanism will theoretically be adopted by all major social media platforms and flag socialized news articles as safe or disputed COVID-19 articles.

1.1. The Role of Machine Learning and Text Mining in Misinformation and Fake News Classification

From the motivation presented in the Introduction section above, it has become clear that computational science in general, and machine learning and computational linguistics, in particular, must be at the forefront of the fight against the infodemic. Prior to the COVID-19 infodemic, machine learning, and natural language processing have played an essential role in fighting misinformation and fake news [19]. We believe that innovating new solutions that leverage the power of both fields is the right step to take in this fight.

The literature is rich of valuable methods and algorithms that demonstrate both the machine learning algorithms and natural language processing approaches, individually or as hybrid. Here we share the background and approaches that represent the backbone of the methods of this paper. In the early 2000′s, Soon et al. claimed that training a machine learning algorithms with specific linguistic features holds a promise in classifying text in general [20]. The authors claimed that their algorithm was the first-learning-based system trained by bigram features to achieve comparable results to non-learning methods.

Mackey et al., in their efforts to identify suspected fake contents on social media, combined natural language processing and machine learning. The approach identified keywords associated with the pandemic and suspected marketing [21]. By analyzing millions of social media posts, the authors adopted a deep learning algorithm that detected high volumes of suspicious and untrustworthy products.

Liu et al., presented a "survey-like" paper to demonstrate the various applications of combining both natural language processing and machine learning. Specifically, this included the method of training algorithms using word features (bigrams) [22]. Bigrams are a sequence of two words that appear in the text (e.g., global pandemic) [23]. They provide valuable and richer textual features than mere single high-frequency word counterparts. Aphiwongsophon et al., demonstrated how famous ML algorithms (e.g., NaiveBayes [24], and Support Vector Machines [25–27] can be used to detect fake news. Their results showed promise with an accuracy of 96% or better [28].

Following a similar path, H. Ahmed et al. also used a classical machine learning algorithm, (i.e., a variation of a support vector machine), but rather trained them using n-gram features [29]. The accuracy of their algorithm was lower than the previous methods (92%). The authors, however, argued that training the algorithm with the n-gram was better in terms of feature quality than features of high frequency that do not contribute to the context of the dataset.

Another interesting approach was employed by Conroy et al. who also used machine learning to detect deception in identifying fake news [30]. The approach combined machine learning, linguistic features (e.g., n-grams), and network analysis for networks of linked data. The authors claim that both linguistic and network analysis methods have shown high accuracy in classification tasks of detecting fake news. The authors concluded their research by making the following recommendations: (1) achieving maximum performance requires deeper linguistic analysis and; (2) the utilization of linked data and a corresponding format will assist in achieving up-to-date fact checking.

1.2. Limitations of Related Studies

The above introduction explains that the related methods motivate the subject and presents the current state of the art. It is clear that both machine learning and text mining present the corner stones for text classification and anomaly detection [31–38]. However, regardless of the underlying algorithmic classification method (naïve bayes, support vector machines), they were all trained from a static set of textual features, such as bigrams. Once the featured were derived, there has been no further work on how the features are related to each other to tell a much bigger story. Our network training model, however, connects the features in the way that the bigrams are naturally connected in the text. This offers the following advantages (1) it makes the model extensible by new datasets without doing the entire training; (2) a network model allows pruning (i.e., getting rid of the noise) using inherent centrality measures (degree, betweenness, closeness, etc.); (3) if necessary, a network model allows multi-label classification by applying network clustering techniques as it has become apparent in the PubMed Case Study below.

1.3. Contributions

The main contributions of this work can be summarized as follows (1) an extensible network model that can be trained with new datasets without retaining and inclusion of previous datasets; (2) the network model enables binary classification if used as it is; (3) the network model can support multi-label classification if it is further analyzed using a network clustering algorithm (e.g., Girvan-Newman algorithm); (4) As stated above, some journal publications have proven to be not credible; the NeoNet algorithm was designed to classify plain text articles. This is further discussed below in the PubMed Case Study section. We have proven that it is possible to train it with publication articles without any changes. Clearly such contributions make the algorithm a general-purpose text classification that may be utilized in various applications (as on social media such as PatientsLikeMe [39] and online medical forums such as Doctor's Lounge [40]).

2. Materials and Methods

With the previous introduction and the recommendations made to fight the COVID-19 infodemic, we present a novel supervised machine learning algorithm, which we call NeoNet. The algorithm was specifically designed for COVID-19 news classification. The overall approach of the NeoNet algorithm is centered around a bigrams network [41]. We applied the Term Frequency Inverse Document Frequency (TF-IDF) algorithm to extract bigrams (a pair of words), which was the bridge to identifying discriminant features. The bigram features naturally present themselves as a network which we used as a training model. Hence, the role of feature selection using TF-IDF to identify bigrams was significant. TF-IDF features have two folds: (1) they provide discriminant features that contribute significantly to the training phase of the algorithm; (2) they provide linked features that take the mere article contents to a connectivity level. The result is an interesting network model that offers a platform for testing whether a new article is relevant from both content and connectivity. In this section, we discuss how the algorithm was designed, implemented, and tested. Particularly, we present the cornerstone steps that lead to determining the class of news articles: (1) textual feature selection is a dominant method, which we used to extract bigram features from news articles [20]; (2) TF-IDF bigram-based network model, which we used for training the algorithm before it was able to predicted the label of new articles, (3) a supervised machine learning algorithm, which predicts the final label for each news article as a safe/disputed COVID-19 news article.

The main dataset used in this work is identified as COVID-19 News Articles Open Research Dataset, which is available at Kaggle [42]. The data existed as a Comma Separated Value (CSV) file that was comprised of seven columns. The ones of interest to this here were the (article title), (article description), and (full article), and it contained 6782 articles. The articles were collected from the website of the Canadian CBC News network. The preprocessing of the text was done using the Pandas [43] framework, and the linguistic

analytics was done using TextBlob [44]. We split the dataset to 10-folds, where each partition contained 500 articles. We used one for training the algorithm and another five to test it. For each test-fold, we set the minimum support parameter to a certain threshold and compared the performance. Figure 1 shows the overall workflow of NeoNet starting with a set of documents (news dataset in this case) and until a classification label is produced.

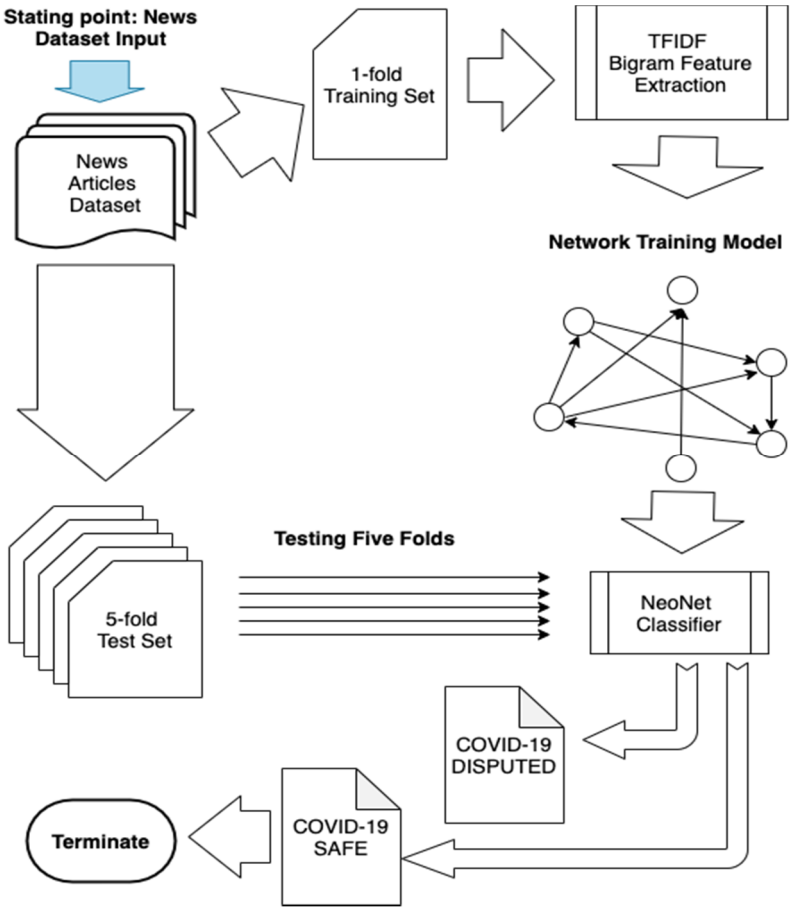

Figure 1. Shows a demonstrative workflow that explains the methods and the various processes starting with a news dataset, analyzing it for bigrams using Term Frequency Inverse Document Frequency (TF-IDF), constructing a network model and training the algorithm with a model to make predictions of 5-fold tests.

2.1. TF-IDF Feature Selection and Model Construction

Every training model starts with a good representation of data items. For text classification in particular, feature selection is the prerequisite step necessary for such a task. Various approaches are designed around the idea of selecting a set of words that best represents the document (or a set of documents). The most common text feature selection that is known is based on the idea of term frequency. Specifically, the Term Frequency and Inverse Document Frequency (TF-IDF) method [45] has been most dominant. In this section, we discuss how we extracted the bigram features needed for training the NeoNet classifier. For this task, we used a COVID-19 news dataset that was trustworthy, and publicly available (published on Kaggle). Due to the fact that raw text presents users with inherent issues (e.g., format, encoding, and punctuation), we performed a preprocessing step to address such issues.

We split the list of articles into 10-folds of 500 articles. We used one-fold to be analyzed for feature selection using the TF-IDF algorithm. Given that a TF-IDF feature can be a word or more, we calibrated the algorithm to capture features that were of exactly two words (i.e., bigram). The TF-IDF scored each feature and ranked them accordingly. When the

TF-IDF was run against the training articles, it produced 193914 bigrams were produced. The TF-IDF measure produces features of a certain confidence. In the training fold (500 articles) of the dataset that we have used 193914 bigrams. Clearly, this caused the model to be noisy which also could have led to an overfitting problem. Therefore, we only selected the top 500 ranked and ignored the rest. Table 1 shows a sample of the top-ranked features selected from the training dataset before the noise removal.

Table 1. Shows a sample of the top features extracted using the TDIDF algorithm.

Order	Feature	Rank	Order	Feature	Rank
3994	COVID-19	20.89461	91,437	world health	2.497454
133,189	public health	6.619656	111,046	new coronavirus	2.469308
30,790	cbc news	5.380869	81,162	https twitter	2.424389
136,126	read happening	4.941779	28,911	care workers	2.418881
97,824	long term	4.706066	76,460	health organization	2.404241
168,712	term care	4.545932	179,072	two weeks	2.394925
129,989	prime minister	4.351588	64,311	federal government	2.304805
76,277	health care	4.169318	159,505	spread covid	2.288158
41,658	coronavirus outbreak	3.968387	76,437	health minister	2.28195
111,316	new york	3.75363	161,592	stay home	2.272373
111,032	new cases	3.704022	151,745	self isolation	2.203682
16,324	around world	3.604401	170,710	the province	2.136745
169,146	tested positive	3.594038	111,509	news network	2.118405
17,315	associated press	3.518788	112,120	non essential	2.051427
124,962	physical distancing	3.507329	59,500	spread coronavirus	1.98592
39,151	confirmed cases	3.472284	86,767	intensive care	1.978987
76,453	health officials	3.272017	123,140	people died	1.977204

Figure 2 is a bar plot that also shed more insight on the ranking and the analysis of the top-40 bigrams. There are four subfigures. The top-left corner contains the top-10 which includes the following bigrams (covid 19, coronavirus outbreak, public health, heath care, New York). The most significant and highly ranked term according to the dataset was “covid 19”. Its corresponding sore according to TF-IDF was 20.9. The rest of the terms fluctuated between 3 and 7, which shows their relevance when compared with the score of the “covid 19” bigram. The remaining bigrams in each of the remaining three subplots contained bigrams such as “tested positive”, “physical distancing”, “coronavirus cases”, “social distancing”, and “federal governments”. Clearly, such bigrams depict an accurate picture of the pandemic in terms of mitigation presented by “social distancing” and “physical distancing”. The global impact of the pandemic was also presented by bigrams such as “federal governments”, “health organizations”, and “public health” concerns.

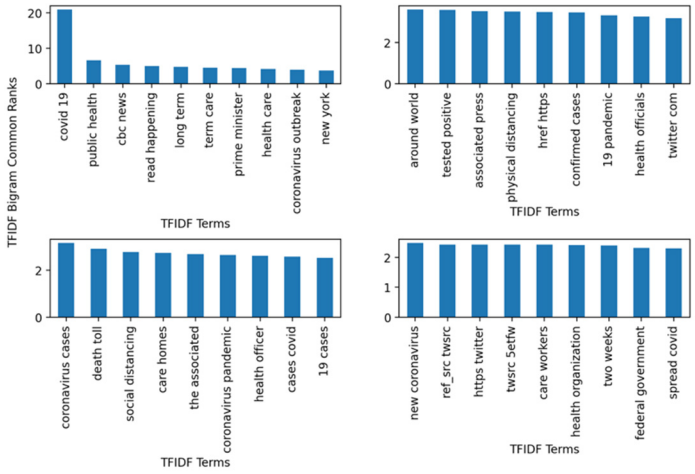

Figure 2. Shows the TF-IDF analysis of top-40 terms in four subfigures. Each figure contains 10 bigrams and their corresponding score.

Bigrams, as a network construction means, are widely used in various computational problems [46–48]. We present an incremental network construction approach that is well-known in the literature on prominent algorithms (e.g., Prim's algorithm [49,50], which starts with an empty set of nodes and incrementally adds new nodes, one node at a time. In a similar fashion, we followed the same method of construction. Our goal was to add all the bigrams that also met a certain criterion. The bigram extraction step, which discussed above, produced the set of length-two features. The length-two not only captures the core necessary features for classification but also offers a network model that can be used for training a classification algorithm. They offer a source-target mechanism where the source and target are nodes in the network and are connected with an edge. The continuation of adding new bigrams forges an incremental linkage. The final outcome of such a process results in a graph where its structure and characteristics are dependent upon the dataset being analyzed (i.e., healthcare, politics, business, etc.). For the COVID-19 domain, following the incremental process ensured an upfront production of high-quality features. The network ensured that classified bigrams were related to the content and not a result of verbatim exact match.

The following example demonstrates how a TF-IDF feature of length-two can provide the foundations for constructing the needed network. A sentence such as "Top U.S. health official Dr. Anthony Fauci said it has a "clear cut, significant, positive effect in diminishing the time to recovery" [51] after favorable results of a clinical trial.", when performing the TF-IDF feature extraction step, produces the following (health official), (clinical trials) bigrams. These two bigrams contribute four different nodes (unigrams), namely, health, official, and clinical, and trials. It would also contribute two edges: an edge from health and official and another from clinical and trials. As we analyze more sentences, we encounter the mention of strict public health measures. In turn, this contributes another bigram (health measure). Putting these bigrams together and connecting them based on the bigram relationship ought to form a graph where the node health is connected to official and measure. The continuous addition of bigrams extracted from the training dataset will result in a much larger network. Figure 3 displays a wordcloud from the top-100 features of the training set. The figure shows proof of how relevant are the features that were selected using the TF-IDF algorithm.

Figure 3. Shows a wordcloud representation of the top-100 features selected using TF-IDF.

Upon constructing the network training model, it ended up with 471 nodes only. This is explained by the fact that some bigrams might share a common word among them. For Example, 'health issues' and 'health problems' have the node 'health' in common. While such bigrams features will be used as they are in other machine learning algorithms, a network model such as ours naturally prunes repeated features that can lead to overfitting which is a problem that other algorithms suffer from. Table 1 summarizes the training model which was constructed from the top-500 TF-IDF features selected, and Figure 4 shows the training model after construction.

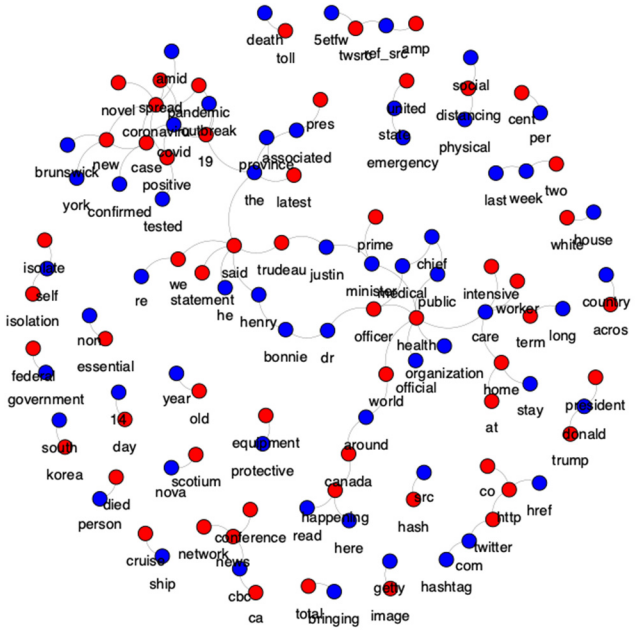

Figure 4. Shows a simplified version of the network training model that is constructed from 100 bigrams features (also produced by TF-IDF).

2.2. The Design of NeoNet Classifier

The previous step explained how a network-based training model was derived from a given set of news articles. Here, we present the algorithmic steps that led to labeling a new article that was yet to be seen by our algorithm. The algorithm was controlled by a configuration parameter, which we called: minimum support, which was inspired by the Apriori algorithm [52–57]. The minimum support guaranteed a certain number of bigrams to be present in each article; otherwise, it would be labeled as suspicious. It ensured that the article contained sufficient contents that contributed to the training model. If the article did not meet this condition it would not be classified as "safe". Clearly, an article that does not have a minimum number of TF-IDF features also does not communicate significant facts that makes it worthy of reading [58]. As for the percentage generated by

the minimum confidence, it guides the setting of the minimum support and helps to set it to a sufficient level. This become significant in long vs short article. In long articles, it was expected to have a higher number of TF-IDF features than shorter articles. If the minimum support parameter was set too low, this percentage helped correct this issue and ensure that news articles were not classified as "safe" if they should be classified as "suspicious". The NeoNet algorithmic steps were described below and also were also expressed in pseudocode in Algorithm 1.

1. Set the minimum support parameter
2. For each new article to be classified
3. Preprocess as previously described
4. Extract TF-IDF bigram features
5. For each component of the bigram (unigram): split into left-unigram and right-unigram
6. Add the left-unigram to the model: if it connects then add the right-bigram.
7. Otherwise, add the right-unigram to the model: if it connects, add the left-unigram and preserve the bigram order
8. Continue until all bigrams are tested
9. Calculate actual support score
10. If support value is above the threshold parameters classify as POSITIVE and assign a COVID19 label
11. Otherwise, classify as NEGATIVE.

Algorithm 1 NeoNet: A Noun-phrase Bigram-based Classification Algorithm

Require: *mini_sup*, the minimum support value v needed
Require: *G*, a graph training model
1: Initilize *min_sup(d)* = *v*, for all $d \in D$
2: Initilize *positive(d)* = 0
3: Initilize *sup_count(d)* = 0
4: **rpt**:
5: **for each** $d \in D$ **do**
6: *bigram_list* ← *extract_bigrams(d)*
7: **for each** *bigram* ∈ *G_list* **do**
8: **if** *left_unigram* ∈ *G* **then**
9: add *G* ← *right_bigram*
10: **else**
11: **if** *right_unigram* ∈ G **then**
12: add *G* ← *right_bigram*
13: **if** *sup_count* ≥ *min_sup*
14: *positive* ← +1
15: *d* ← *COVID*19 label
16: **until** no more document to classify

Algorithm 1 shows the steps of the NeoNet algorithm in pseudocode staring from when the bigram features were extracted until a classification label was generated.

3. Experiments and Results

Using the training model resulting from the bigram feature selection step, we conducted a series of classification (testing) experiments. Using five different folds of the dataset, and different configurations of the minimum support parameter, we measured the precision of the NeoNet algorithm. The rationale behind this was to come up with a threshold that produced the best outcome. The minimum support parameter was based on the number of bigrams produced by each article. The higher the number of the bigrams matching, the higher the chances of an article being classified as a positive COVID-19 class. However, the experimentation guided the algorithm to identify a reasonable threshold. A very high number of bigrams would lead to classifying articles that were extremely similar in content. As a result, the algorithm would miss articles that belonged to the COVID-19 class, but less similar to the training model. On the other hand, a very low threshold would lead to classifying any article with a slight overlap as positive and would have made the algorithm not precise. We used the following min-sup configurations [5, 10, 15, 20, 25] bigrams. Figure 5 shows the five different test sets and how they were classified according

to NeoNet, with various minimum support levels. The analysis shows that mandating that at least 5 bigrams were matched against the training set and produced a precision value between 99.6%–100%. If it was set to a more demanding parameter (mini-sup = 25) the classification results fluctuate between 91.18%–95.59%. The experimentations showed that somewhere in the middle is reasonable (min-sup = 15), and it produced a classification precision that fluctuated between 97.99%–99.20%. Each fold was plotted using five different curves, one for each minimum support value.

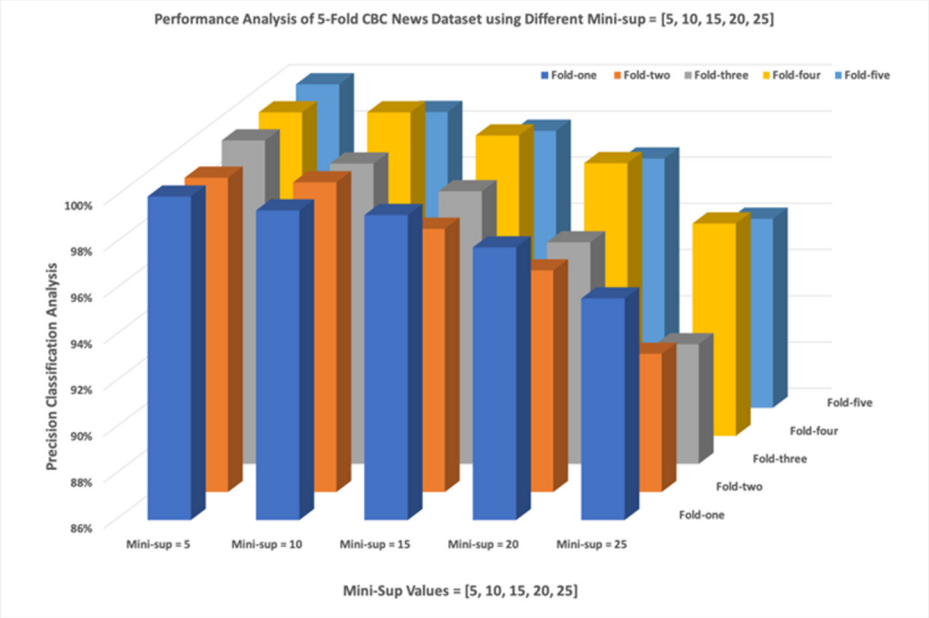

Figure 5. Shows the performance of NeoNet with different configurations of the min-sup parameter against CBC News Dataset using min-sup levels: [5, 10, 15, 20, 25].

We have shown above how NeoNet could be controlled using the min support to make it flexible to use in various case scenarios. However, the algorithm also performs exceptionally well without such configurations when needed. In this section we show its performance analysis compared with the most prominent algorithms (e.g., SVM, neural nets, random forests). Given the fact that such algorithms do not necessarily utilize a similar notion of the min support/confidence, we set both parameters to min-sup = 15. The algorithm was trained using a 500 articles fold. The rest of the dataset was split into 5-folds, each of 500 articles.

As expected, each fold of the dataset was tested against NeoNet and compared with a counterpart algorithm. The algorithm was tested on each of the five folds and compared against all other algorithms to measure the precision achieved for all of the 5-folds. Figure 6 below shows how NeoNet's performance (shown on the far left of the x-axis) outperformed all other algorithms, unless a perfect classification results were achieved by both algorithms (e.g., NeoNet vs. Neural Net). The diagram shows a common theme: Except for NeoNet, Fold-3 (depicted in grey) appeared to be scored the lowest among all the algorithms. Another noteworthy observation is that Fold-1 and Fold-5 also appeared to be scored the highest (99.2%) among several algorithms which include Stochastic Gradient Descend, SVM, Random Forests, and Neural Nets. This was as close as it got when their precision was compared with NeoNet. However, when all of the algorithms failed to achieve a perfect precision, NeoNet showed dominance and outperformed all algorithms. We conducted the experiment using the Orange Data Mining Toolbox in Python [59].

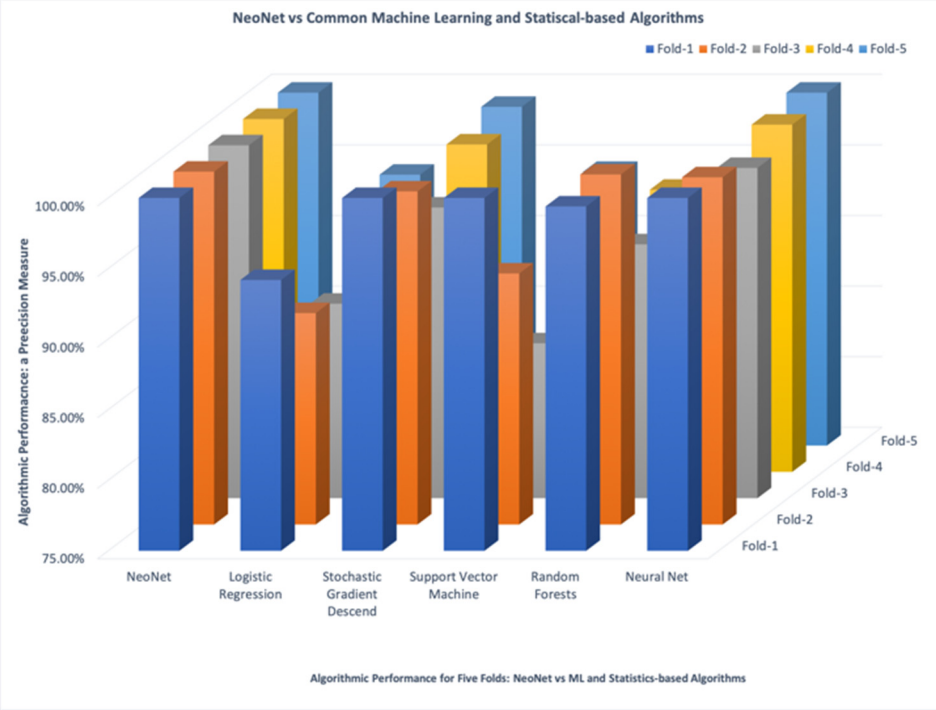

Figure 6. Shows the performance analysis of NeoNet vs. the most prominent Machine Learning algorithms and other common statistical-based methods.

A reasonable explanation for the outstanding performance that NeoNet demonstrated is the feature selection part when training the algorithm. Clearly, the TF-IDF bigram features were more indicative than counterpart unigrams with mere frequency. Bigrams such as (covid, '19'), (coronavirus, 'outbreak'), (social, 'distancing'), ('physical', 'distancing'), (self, 'isolation'), ('coronavirus', 'pandemic'), to list a few, were extremely indicative of the domain being analyzed. Additionally, the training model that was used is a natural extension of the individual bigram features, because it naturally forms a network that accepted terms in common and rejected terms that don't contribute to the overall model. On the other hand, all prominent algorithms were driven from a model that relied on the individual terms and their frequencies. They all failed to integrate the associations with other terms. Clearly, such integration of these two characteristics (TF-IDF bigram features and a network model) indeed contributed to an exceptional performance demonstrated by the NeoNet algorithm that has been presented here.

The PubMed Case Study

The Experiment section discussed above has shown the new methods and approaches this research have taken to produce a label for never-before-seen news articles. However, the premise of the algorithm was to show that it could work the same way with other textual inputs, such as medical publications, doctors' notes, etc. In this section we demonstrate another case scenario that will show how NeoNet is a text-classification general purpose algorithm that can perform the same way, regardless of the input source type (news, online medical forums, doctors' notes, or medical publication). Starting with a publication dataset extracted from PubMed, after searching the web portal for keywords such as (COVID-19 and coronavirus). Although, the search results produced more than 2500 medical abstracts (approx. 100, 000), we only used 500 abstracts for training, and 5-fold of 500 abstracts for testing. This was indeed consistent with the same experiments performed against the CBC news dataset.

Following the same processes explained above and also demonstrated in Figure 1, we extracted the TF-IDF bigrams and constructed a similar training network model. The

generated network constructed had the following characteristics: (1) number of nodes: 467, (2) number of edges: 330, and (3) average degree: 2.83. Table 2 below draws a comparison between the two training models generated from the two datasets. While the two datasets produced a relatively close number of nodes, the PubMed dataset had a significantly smaller number of edges. This is explained by the fact that publications covered various "clusters" of public health issues, such as the vaccine development, drug treatment, the COVID-19 disease, and its impact on the human body, among other things. On the other hand, news articles addressed the general public in much less domain-specific but commonly related terms.

Table 2. Describes the structural properties of the network training models generated from two different datasets.

DataSet	Nodes	Edges	AVG Degree
CBC News	412	471	2.28
PubMed	467	330	2.83

As for the actual classification results, we performed similar experiments against the 5-fold test sets provided by the PubMed dataset. Each fold was tested using a minimum support of the following values [5, 10, 15, 20, 25]. We observed a very similar pattern: the more the number of bigrams needed to classify a document as a COVID-19 article, the lower the precision of the classification. For example, when the classification required 5-bigrams (i.e., 10 connected terms in the training model) the classification precision fluctuated between 98–99.4%. In another case, when the minimum support was set to 25-bigrams (i.e., 50 connected terms in the training model), the classification precision fluctuated between 89.98–91.98%.

Figure 7 shows the entire analysis of each fold and precision resulting from the various minimum support configurations. When comparing the precision derived from the 5-fold CBC News dataset, we found that precision fluctuated between 99.0–100% when the minimum support was set to 5-bigrams. When the minimum support was set to 25-bigrams, the precision fluctuated between 94.18–95.59%. Clearly, the drop-in precision in the PubMed 5-folds was due the training model being less connected, due to the significantly lower number of edges in its training model. We believe the precision could be enhanced in the case if multilabel classification (vaccine development, drug treatment, and COVID-19 disease and symptoms, etc.) This finding requires further investigation in the future to assess how the training model can offer more insights using the underlying inherent subtopics.

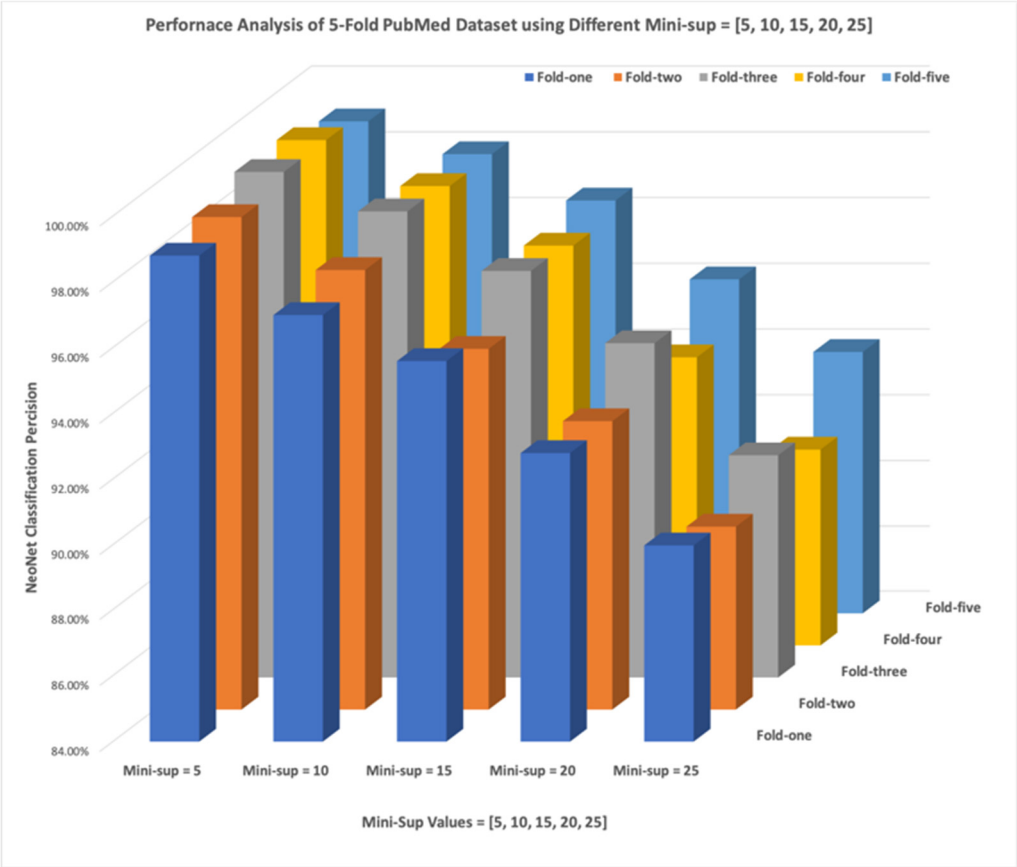

Figure 7. Shows the NeoNet precision with different configurations of the min-sup parameter while classifying a PubMed Coronavirus Dataset. The figure shows five different configurations of the min-sup [5, 10, 15, 20, 25]. The plots are displayed from left to right respective to the values of the configurations.

4. Discussions & Future Directions

In this article, we discussed the how the COVID-19 pandemic has also been accompanied but an infodemic. Particularly, we discussed the various aspects of the infodemic and how it presents a serious health threat to the general public due to the misinformation/disinformation that may exist in the source (e.g., scientific publications, fake news, and social media posts). For instance, we presented evidence of disinformation that existed in a publication, which eventually was presumed to be from a suspicious source. The article reported health issues associated with vitamin D. As the article was published, it was also highlighted by a reputable news organization (i.e., DailyMail). The matter was made worse when the DailyMail news article [60] was also socialized on Facebook and Twitter. Clearly, such misinformation or (disinformation in this case) threaten the world's public health.

This paper also highlighted the various efforts that have been taken by the scientific community in the fight against the infodemic and made recommendations. One specific reference, Eysenbach, addressed the infodemic and introduced four pillars that must be observed in order to win this fight. The recommendations included information monitoring and encouraging knowledge refinement and quality improvement processes. Our research here has taken such recommendations into serious consideration and implemented them accordingly. Specifically, we presented an information monitoring and a knowledge refinement solution that addressed the problem from the source. The research also performed a diligent literature review on what specific tools and research methods should be used. The technical recommendations were influenced by advances of machine learning, computational linguistic, and network science. Indeed, this paper has presented a novel machine

learning algorithm that utilized knowledge refinement produced by natural language processing to produce training features. We then empowered the algorithm by a network model. Such a model offered both the structural components (i.e., nodes and edges) and the node degree centrality to perform the knowledge refinement when constructing the training model. This led to the generation of highly representative features and eliminated the noise by using the degree centrality as a heuristic.

As for the actual step of training and testing the algorithm, we selected a trustworthy set of news articles which was published on Kaggle and divided it into five-folds. We performed five different experiments to come up with a reasonable min-sup threshold. Each experiment was performed against the 5-folds with a given configuration of the min-support. The experiment was repeated with 5, 10, 15, 20, and 25, and it showed that a threshold of 15 produced the best results without being too strict or too noisy. This specific threshold produced a classification precision that fluctuated from 97.99–99.20%. Such results are indeed promising as the algorithm selected relevant and high-quality features that represent the main content of any domain. The minimum support parameter makes the algorithm flexible for the domain experts to experiment with various dataset of different characteristics which helps to achieve the best classification results. The flexibility of tuning the minimum support parameter makes the NeoNet viable and adaptable to various situations. It can be set aggressively in situations where suspicious sources are common, while it can be relaxed in the case of more reliable news organizations. By testing the algorithm on a COVID-19 dataset, we believe that we directly contributed to the pillars of fighting the infodemic and have indeed shown how the algorithm conquers misinformation/disinformation propagated on the web in the form of news articles. The introduction of a network approach, which was based on TF-IDF features for training its model, has taken the text classification from a mere content matching level to a connectivity level expressed by the underlying relationships that make up the training model. The future direction of this work will consider developing an adaptive approach to set such a configuration automatically. We will also consider promoting the algorithm to be multi-lingual and test it against various datasets from various news organizations.

It is worthwhile mentioning that the algorithm would perfectly function on all other text sources, not only the news. As demonstrated in the PubMed Case Study, we also expect the algorithm to function the exact same way, and without any modifications, to online medical forums or doctor's notes. The setting of the minimum support parameter will require calibration by experimentation. Our reason to believe that NeoNet will be successful is that it was already tested on two different types of data (news articles and medical publications) and produced comparable results. That is due to the fact that the algorithm was trained using bigram features extracted from full-text. Such features are highly significant in the context of medical publications since they may reference entities such as organs, disease, gene, protein, indication, symptoms, etc. Eventually, the training model will be rich, and suspected sources will fail to be classified positively against the model. We also believe that adding features from doctor's notes to the training model of the scientific literature will eliminate suspicious ones, such as false the reference that promoted vitamin D. This is yet to be explored in future publications.

Testing the algorithm against an entire set of automatically generated fake news is another interesting future direction the authors are considering. By using means of automatic generation of text, we used one of the same datasets to generate such text. Such a test will provide future merits on how to continue to fight the COVID-19 infodemic.

5. Conclusions

To conclude, we presented NeoNet, a general-purpose supervised machine learning algorithm that analyzes textual content and produces an extensible network training model. The purpose of this algorithm was to flag textual contents as COVID-19 safe or disputed for the general public to read. The algorithm was demonstrated and tested against two different datasets of different natures: (1) a publicly available news dataset (CBC News)

and (2) COVID-19 medical publications publicly available from PubMed. NeoNet showed promise in the battle against the COVID-19 infodemic when compared against prominent machine learning algorithms (Support Vector Machines, Artificial Neural Networks, and Decision Trees). Additionally, NeoNet outperformed other statistical-based methods such as Stochastic Gradient Descend and Logistic Regression.

The method and the analysis of this research indeed confirmed the invasion and widespread of misinformation in major news organizations (e.g., DailyMail), social media, and even academic publishers. Our research highlighted the dire need and the importance of adopting new defensive mechanism and procedures against such misinformation and other forms of disputed digital contents. While prominent social media platforms (e.g., Twitter) have taken initial steps towards flagged disputable contents from influential figures, such issues remain desperate and urgent. The findings of our research beg for adopting new digital publishing models and an overall digital transformation movement that guarantees the freedom of speech and credibility. As demonstrated above, our algorithm provides an intelligent tool for all online text-producing organizations. Giving the promise demonstrated by the analysis of this work, we believe that such a tool provides a deeper defensive mechanism than existing counterparts when new policies are made and are ready to be adopted.

Author Contributions: M.A.R.A.: The Principle Investigator and the one who secured the initial funding of this research. The proposal ideas were the core components which have translated into research components. He guided the research directions through the various phases. A.A.H.: The NeoNet algorithm designer, the hands-on scientists who have implemented the algorithm, tested it, suggested new directions to the PI and the senior author, the one performed all evaluations during the lifecycle of this work. He is the main manuscript writer and the one who was in charge of communication either with the team or the respected reviewers. X.W.: The senior author of this work. He has guided the overall research directions and have provided significant insights during the various phases of algorithm evaluation. X.W. has also provided directions on how to respond to the reviewers in a satisfactory matter. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partly funded by IU of Madinah, Tamayoz initiative project 23/40 and National Security Agency #22341 Cyber Institute.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors thank Zuzana Mikulecká for the valuable discussions around the use of new datasets, case studies, and during the process of responding to various points raised by the respected reviewers. The authors also thank Regis O'Connor for providing her expertise producing some of the figures of this work. Lastly, the authors thank Eszter Szenes for the valuable discussions around multilingual contents and future feature selections to consider for the classification.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Cuan-Baltazar, J.Y.; Muñoz-Perez, M.J.; Robledo-Vega, C.; Pérez-Zepeda, M.F.; Soto-Vega, E. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health Surveill.* **2020**, *6*, e18444. [CrossRef] [PubMed]
2. Hou, Z.; Du, F.; Zhou, X.; Jiang, H.; Martin, S.; Larson, H.; Lin, L. Cross-Country Comparison of Public Awareness, Rumors, and Behavioral Responses to the COVID-19 Epidemic: Infodemiology Study. *J. Med. Internet Res.* **2020**, *22*, e21143. [CrossRef] [PubMed]
3. Moon, H.; Lee, G.H. Evaluation of Korean-Language COVID-19–Related Medical Information on YouTube: Cross-Sectional Infodemiology Study. *J. Med. Internet Res.* **2020**, *22*, e20775. [CrossRef] [PubMed]
4. Rovetta, A.; Bhagavathula, A.S. Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. *J. Med. Internet Res.* **2020**, *22*, e20673. [CrossRef] [PubMed]
5. Tang, N.; Bai, H.; Chen, X.; Gong, J.; Li, D.; Sun, Z. Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. *J. Thromb. Haemost.* **2020**, *18*, 1094–1099. [CrossRef]

6. Tangcharoensathien, V.; Calleja, N.; Nguyen, T.; Purnat, T.; D'Agostino, M.; Garcia-Saiso, S.; Landry, M.; Rashidian, A.; Hamilton, C.; AbdAllah, A.; et al. Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation. *J. Med. Internet Res.* **2020**, *22*, e19659. [CrossRef]
7. Gazendam, A.; Ekhtiari, S.; Wong, E.; Madden, K.; Naji, L.; Phillips, M.; Mundi, R.; Bhandari, M. The "Infodemic" of Journal Publication Associated with the Novel Coronavirus Disease. *J. Bone Joint Surg. Am.* **2020**, *102*, e64. [CrossRef] [PubMed]
8. Okan, O.; Bollweg, T.M.; Berens, E.M.; Hurrelmann, K.; Bauer, U.; Schaeffer, D. Coronavirus-related health literacy: A cross-sectional study in adults during the COVID-19 infodemic in Germany. *Int. J. Environ. Res. Public. Health* **2020**, *17*, 5503. [CrossRef] [PubMed]
9. Morley, J.; Cowls, J.; Taddeo, M.; Floridi, L. Public Health in the Information Age: Recognizing the Infosphere as a Social Determinant of Health. *J. Med. Internet Res.* **2020**, *22*, e19311. [CrossRef]
10. Dong, W.; Tao, J.; Xia, X.; Ye, L.; Xu, H.; Jiang, P.; Liu, Y. Public Emotions and Rumors Spread During the COVID-19 Epidemic in China: Web-Based Correlation Study. *J. Med. Internet Res.* **2020**, *22*, e21933. [CrossRef] [PubMed]
11. Stephens, M. A geospatial infodemic: Mapping Twitter conspiracy theories of COVID-19. *Dialogues Hum. Geogr.* **2020**, *10*, 276–281. [CrossRef]
12. Islam, M.S.; Sarkar, T.; Khan, S.H.; Kamal, A.-H.M.; Hasan, S.M.M.; Kabir, A.; Yeasmin, D.; Islam, M.A.; Chowdhury, K.I.A.; Anwar, K.S.; et al. COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *Am. J. Trop. Med. Hyg.* **2020**, *103*, 1621–1629. [CrossRef]
13. Orso, D.; Federici, N.; Copetti, R.; Vetrugno, L.; Bove, T. Infodemic and the spread of fake news in the COVID-19-era. *Eur. J. Emerg. Med.* **2020**, *27*, 327–328. [CrossRef]
14. Matthews, S. Government Orders Review into Vitamin D Role in Covid-19. Available online: https://www.dailymail.co.uk/news/article-8432321/Government-orders-review-vitamin-D-role-Covid-19.html (accessed on 4 August 2021).
15. Bunker, D. Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic. *Int. J. Inf. Manag.* **2020**, *55*, 102201. [CrossRef] [PubMed]
16. Eysenbach, G. How to Fight an Infodemic: The Four Pillars of Infodemic Management. *J. Med. Internet Res.* **2020**, *22*, e21820. [CrossRef] [PubMed]
17. Gallotti, R.; Valle, F.; Castaldo, N.; Sacco, P.; De Domenico, M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat. Hum. Behav.* **2020**, *4*, 1285–1293. [CrossRef]
18. Twitter to Start Removing COVID-19 Vaccine Misinformation. Available online: https://apnews.com/article/misinformation-immunizations-coronavirus-pandemic-085cc1b49a5d488026f2e59d8f32d590 (accessed on 25 December 2020).
19. Braşoveanu, A.M.P.; Andonie, R. Semantic Fake News Detection: A Machine Learning Perspective. In Proceedings of the Advances in Computational Intelligence, Gran Canaria, Spain, 12–14 June 2019; Rojas, I., Joya, G., Catala, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 656–667.
20. Soon, W.M.; Ng, H.T.; Lim, D.C.Y. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Comput. Linguist.* **2001**, *27*, 521–544. [CrossRef]
21. Mackey, T.K.; Li, J.; Purushothaman, V.; Nali, M.; Shah, N.; Bardier, C.; Cai, M.; Liang, B. Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram. *JMIR Public Health Surveill.* **2020**, *6*, e20794. [CrossRef] [PubMed]
22. Liu, F.; Liu, F.; Liu, Y. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In Proceedings of the 2008 IEEE Spoken Language Technology Workshop, Goa, India, 15–19 December 2008; pp. 181–184.
23. Relationships between Words: N-Grams and Correlations—Text Mining with R [Book]. Available online: https://www.oreilly.com/library/view/text-mining-with/9781491981641/ch04.html (accessed on 25 December 2020).
24. Qiang, G. An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification. In Proceedings of the 2010 Second International Conference on Computer Research and Development, Kuala Lumpur, Malaysia, 7–10 May 2010; ISBN 978-0-7695-4043-6.
25. Meyer, D.; Leisch, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55*, 169–186. [CrossRef]
26. Suthaharan, S. Support Vector Machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*; Suthaharan, S., Ed.; Integrated Series in Information Systems; Springer: Boston, MA, USA, 2016; pp. 207–235, ISBN 978-1-4899-7641-3.
27. What Is a Support Vector Machine? | Nature Biotechnology. Available online: https://www.nature.com/articles/nbt1206-1565 (accessed on 25 December 2020).
28. Aphiwongsophon, S.; Chongstitvatana, P. Detecting Fake News with Machine Learning Method. In Proceedings of the 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Rai, Thailand, 18–21 July 2018; pp. 528–531.
29. Ahmed, H.; Traore, I.; Saad, S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In Proceedings of the Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Vancouver, BC, Canada, 25–27 October 2017; Traore, I., Woungang, I., Awad, A., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 127–138.
30. Conroy, N.K.; Rubin, V.L.; Chen, Y. Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–4. [CrossRef]

31. Dunning, T.; Friedman, E. *Practical Machine Learning: A New Look at Anomaly Detection*; O'Reilly Media, Inc.: Newton, MA, USA, 2014; ISBN 978-1-4919-1418-2.
32. Inoue, J.; Yamagata, Y.; Chen, Y.; Poskitt, C.M.; Sun, J. Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning. In Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 1058–1065.
33. Kang, D.-K.; Fuller, D.; Honavar, V. Learning classifiers for misuse and anomaly detection using a bag of system calls representation. In Proceedings of the Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop, West Point, NY, USA, 15–17 June 2005; pp. 118–125.
34. Liu, D.; Zhao, Y.; Xu, H.; Sun, Y.; Pei, D.; Luo, J.; Jing, X.; Feng, M. Opprentice: Towards Practical and Automatic Anomaly Detection Through Machine Learning. In Proceedings of the 2015 Internet Measurement Conference, Tokyo, Japan, 28–30 October 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 211–224.
35. Omar, S.; Ngadi, M.; Jebur, H.; Benqdara, S. Machine Learning Techniques for Anomaly Detection: An Overview. *Int. J. Comput. Appl.* **2013**, *79*. [CrossRef]
36. Pecht, M.G.; Kang, M. Machine Learning: Anomaly Detection. In *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*; IEEE: New York, NY, USA, 2019; pp. 131–162, ISBN 978-1-119-51530-2.
37. Shon, T.; Moon, J. A hybrid machine learning approach to network anomaly detection. *Inf. Sci.* **2007**, *177*, 3799–3821. [CrossRef]
38. Shon, T.; Kim, Y.; Lee, C.; Moon, J. A machine learning framework for network anomaly detection using SVM and GA. In Proceedings of the Sixth Annual IEEE SMC Information Assurance Workshop, West Point, NY, USA, 15–17 June 2005; pp. 176–183.
39. PatientsLikeMe. Available online: https://www.patientslikeme.com/ (accessed on 10 July 2021).
40. Medical News, Opinion, Health Information, Journal and Conference Briefings, Industry Alerts on Doctors Lounge. Available online: https://www.doctorslounge.com/ (accessed on 10 July 2021).
41. Hamed, A.A.; Ayer, A.A.; Clark, E.M.; Irons, E.A.; Taylor, G.T.; Zia, A. Measuring climate change on Twitter using Google's algorithm: Perception and events. *Int. J. Web Inf. Syst.* **2015**, *11*, 527–544. [CrossRef]
42. COVID-19 Open Research Dataset Challenge (CORD-19). Available online: https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (accessed on 25 December 2020).
43. Pandas—Python Data Analysis Library. Available online: https://pandas.pydata.org/ (accessed on 25 December 2020).
44. TextBlob—Google Search. Available online: https://www.google.com/search?q=TextBlob&oq=TextBlob&aqs=chrome..69i57j35i39j69i59j0l5.2340j0j4&sourceid=chrome&ie=UTF-8 (accessed on 8 July 2021).
45. Ramos, J. Using TF-IDF to Determine Word Relevance in Document Queries. pp. 1–4. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf (accessed on 29 July 2021).
46. Bekkerman, R.; Allan, J. *Using Bigrams in Text Categorization*; Center of Intelligent Information Retrieval, UMass Amherst: Amherst, MA, USA, 2004; pp. 1–10.
47. Hachaj, T.; Ogiela, M.R. What Can Be Learned from Bigrams Analysis of Messages in Social Network? In Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 13–15 October 2018; pp. 1–4.
48. Tan, C.-M.; Wang, Y.-F.; Lee, C.-D. The use of bigrams to enhance text categorization. *Inf. Process. Manag.* **2002**, *38*, 529–546. [CrossRef]
49. Dey, A.; Pal, A. Prim's algorithm for solving minimum spanning tree problem in fuzzy environment. *Ann. Fuzzy Math. Inform.* **2016**, *12*, 419–430.
50. Wang, W.; Huang, Y.; Guo, S. Design and Implementation of GPU-Based Prim's Algorithm. *Int. J. Mod. Educ. Comput. Sci.* **2011**, *3*, 55–62. [CrossRef]
51. CBC News. The Latest on the Coronavirus Outbreak for May 1 | CBC News. Available online: https://www.cbc.ca/news/the-latest-on-the-coronavirus-outbreak-for-may-1-1.5552899 (accessed on 4 August 2021).
52. Al-Maolegi, M.; Arkok, B. An Improved Apriori Algorithm for Association Rules. *arXiv Prepr.* **2014**, arXiv:14033948.
53. Li, N.; Zeng, L.; He, Q.; Shi, Z. Parallel Implementation of Apriori Algorithm Based on MapReduce. In Proceedings of the 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Kyoto, Japan, 8–10 August 2012; pp. 236–241.
54. Perego, R.; Orlando, S.; Palmerini, P. Enhancing the Apriori Algorithm for Frequent Set Counting. In Proceedings of the Data Warehousing and Knowledge Discovery, Munich, Germany, 5–7 September 2001; Kambayashi, Y., Winiwarter, W., Arikawa, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 71–82.
55. Singh, J.; Ram, H.; Sodhi, D.J.S. Improving Efficiency of Apriori Algorithm Using Transaction Reduction. *Int. J. Sci. Res. Publ.* **2013**, *3*, 1–4.
56. Toivonen, H. Apriori Algorithm. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2010; pp. 39–40, ISBN 978-0-387-30164-8.
57. Ye, Y.; Chiang, C.C. A Parallel Apriori Algorithm for Frequent Itemsets Mining. In Proceedings of the Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06), Seattle, WA, USA, 9–11 August 2006; pp. 87–94.

58. Alonso-Reina, A.; Sepúlveda-Torres, R.; Saquete, E.; Palomar, M.; Team GPLSI. Approach for automated fact checking. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), Hong Kong, China, 3 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 110–114.
59. Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
60. "Alarmingly High" Proportion of British People Are Vitamin D Deficient | Daily Mail Online. Available online: https://www.dailymail.co.uk/sciencetech/article-9068299/Alarmingly-high-proportion-British-people-vitamin-D-deficient.html (accessed on 25 December 2020).