 doi:10.20944/preprints202106.0482.v1

Article

Fighting the COVID-19 Infodemic with NeoNet: A Text-based Supervised Machine Learning Algorithm

Mohammad AR Abdeen[1], Ahmed Abdeen Hamed[2,*] and Xindong Wu[3]

1 Faculty of Computer and Information Systems; Islamic University of Madina; mabdeen@iu.edu.sa
2 School of Cybersecurity, Data Science and Computing; Norwich University; ahamed@norwich.edu
3 Mininglamp Technology; Mininglamp Academy of Science; wuxindong@mininglamp.com

* Correspondence: ahamed@norwich.edu; Tel.: +1 812 360 2703

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 and follow proper procedures to mitigate the risks. Here, we present a novel supervised machine learning text mining algorithm that analyzes the content of a given news article and assign a label to it. The NeoNet algorithm is trained by noun-phrases features which contributes a network model. The algorithm was tested on a real-world dataset and predicted the label of never-seem articles and flags ones that are suspicious or disputed. In five different fold comparisons, NeoNet surpassed prominent contemporary algorithm such as Neural Networks, SVM, and Random Forests. The analysis shows that the NeoNet algorithm predicts a label of an article with a 100% precision using a non-pruned model. This highlights the promise of detecting disputed online contents that may contribute negatively to the COVID-19 pandemic. Indeed, using machine learning combined with powerful text mining and network science provide the necessary tools to counter the spread of misinformation, disinformation, fake news, rumors, and conspiracy theories that is associated with the COVID19 Infodemic.

Keywords: COVID-19 Infodemic, Text Classification, Noun-phrases Networks, Supervised Learning.

1. Introduction

Without doubt, the Coronavirus pandemic has affected the world around us in unprecedented way. 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 [1–7]. According to the World Health Organization, an infodemic is "an overabundance of information – some accurate and some not."[8]. 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 has not gone through rigorous peer-review process [9]. 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 [10]. 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 [11]

The coronavirus infodemic aspects are many: (1) The spread of rumors across the world has led to inappropriate behavior and have caused adverse effect on people's physical and the mental health [2,12] (2) conspiracy theories have widespread during the pandemic in attempt to explain the unusual circumstances [13]. In fact, similar theories have emerged during the SARS outbreak in China and Ebola outbreak the Congo [14]. (3)

 doi:10.20944/preprints202106.0482.v1

misinformation and public health damage 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 citizen of his country not to use certain drugs (e.g., cortisone); an advice has gone viral during the pandemic. 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 [15]. (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) [14], (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 was never linked to the listed affiliation. Such an article was downloaded 17000 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 as: "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" [16]. 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 is answered an article titled: "Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic" [17]. 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 titled: "How to fight an Infodemic: The Four Pillars of Infodemic Management" [18]. 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, (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 [8]. The effort resulted in recommendations that also overlapped with the four pillars presented in [18] (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 [19], 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 its 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 become viral. An essential 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

 doi:10.20944/preprints202106.0482.v1

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 [20]. In the deal scenario, we envision that our mechanism is theoretically 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 the COVID-19 infodemic, machine learning, and natural language processing have played an essential role in fighting misinformation and fake news [21]. 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. The authors claimed that their algorithm is the first-learning-based system trained by bigram features to achieve comparable results to non-learning methods [22].

Mackey et al., in their efforts of identifying suspected fake contents on social media, they combined natural language processing and machine learning. The approach identified keywords associated with the pandemic and suspected marketing [23]. 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, the method of training algorithms using word features (bigrams) [24]. Bigrams are a sequence of two words that appear in the text (e.g., global pandemic) [25]. They provide valuable and richer textual features than mere single high-frequency words counterparts. Aphiwongsophon et al., demonstrated how famous ML algorithms (e.g., NaiveBayes [26,27], and Support Vector Machines [28–30]) can be used to detect fake news. Their results shown promise with accuracy of 96% or better [31].

Following a similar path, H. Ahmed et al., also used classical machine learning algorithms, (i.e., a variation of support vector machine), but rather trained them using n-gram features [32]. The accuracy of their algorithm was lower than the previous methods (92%). The authors, however, argued that training the algorithm with the n-gram is better in terms of feature quality than features of high frequency that do not contribute to the context of the dataset.

Another interesting approach but Conroy et al., who also used machine learning to detect deception in identifying fake news. 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 conclude their research by making the following recommendations: (1) achieving maximum performance require

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

deeper linguistic analysis of, (2) the utilization of linked data and corresponding format will assist in achieving up-to-date fact checking [33].

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 is specifically designed for COVID-19 news classification. The overall approach of the NeoNet algorithm is centered around a bigrams network. We applied the TFIDF algorithm [34] to extract bigrams (a pair of words) which is the bridge to identifying discriminant features. The bigram features naturally present themselves as a network which we use as a training model. Hence, the role of feature selection using TFIDF to identify bigrams is significant. TFIDF features have two folds: (1) provide discriminant features that contribute significantly to the training phase of the algorithm, (2) 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 is designed, implemented, and tested. Particularly, we present the cornerstone steps that lead to determining the class of news articles: (1) TFIDF feature selection, which is used to extract bigram features from news articles [23,35] (2) TFIDF bigram-based network model, which we use for training the algorithm before it is able to predict 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 dataset used in this work is identified as (COVID-19 News Articles Open Research Dataset) which is available at Kaggle [36]. The data exists as a Comma Separated Value (CSV) file that is comprised of seven columns. The ones of interest to this here are the (article title), (article description), and (full article) and contains 6782 articles. The articles collected from the website of the Canadian CBC News network. The preprocessing of the text is done using the Pandas [37] framework and the linguistic analytic is done using TextBlob [38]. We split the dataset 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.

2.1 TFIDF 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 known is based on the idea of term frequency. Specifically, the Term Frequency and Inverse Document Frequency (TFIDF) method [32,34] 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 is trustworthy, and publicly available (published on Kaggle [36]). 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 TFIDF algorithm. Given that a TFIDF feature can be a word or more, we calibrated the algorithm to capture features that are of exactly 2 words (i.e., bigram). The TFIDF scores each feature and rank them accordingly. When the TFIDF was run against the training articles it produced 193914 bigrams. The TFIDF measure produces features of a certain confidence. In the training fold (500 articles) of the dataset that we have used produced 193914 bigrams. Clearly, this causes the model to be noise and also could lead to an overfitting problem. Therefore, we only selected the top 500 ranked and ignored the rest. Table 1 shows a sample of top-ranked features selected from the training dataset before the noise removal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

Order	Feature	Rank	Order	Feature	Rank
3994	covid 19	20.89461	**91437**	world health	2.497454
133189	public health	6.619656	**111046**	new coronavirus	2.469308
30790	cbc news	5.380869	**81162**	https twitter	2.424389
136126	read happening	4.941779	**28911**	care workers	2.418881
97824	long term	4.706066	**76460**	health organization	2.404241
168712	term care	4.545932	**179072**	two weeks	2.394925
129989	prime minister	4.351588	**64311**	federal government	2.304805
76277	health care	4.169318	**159505**	spread covid	2.288158
41658	coronavirus outbreak	3.968387	**76437**	health minister	2.28195
111316	new york	3.75363	**161592**	stay home	2.272373
111032	new cases	3.704022	**151745**	self isolation	2.203682
16324	around world	3.604401	**170710**	the province	2.136745
169146	tested positive	3.594038	**111509**	news network	2.118405
17315	associated press	3.518788	**112120**	non essential	2.051427
124962	physical distancing	3.507329	**59500**	spread coronavirus	1.98592
39151	confirmed cases	3.472284	**86767**	intensive care	1.978987
76453	health officials	3.272017	**123140**	people died	1.977204

Table 1 shows a sample of top features extracted using the TDIDF algorithm. The first column shows its order in the data, the second column shows the actual feature being selected, and the last column displays the rank of the feature in the dataset.

Bigrams, as network construction means, have been widely used in various computational problems [39]. We present an incremental network construction approach that is well-known in the literature in prominent algorithms (e.g., Prim's algorithm [40,41]) which starts with an empty set of nodes and incrementally adds new nodes, one node at a time. In a similar fashion, we follow the same method of construction. Our goal is to add all the bigrams that also meet 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 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 ensures an upfront production of high-quality features. The network ensures that classified bigrams are related to the content and not a result of verbatim exact match.

The following example demonstrates how a TFIDF 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" [42] after favorable results of a clinical trial.) When performing the TFIDF feature extraction step, it produces the following (health official), (clinical trials) bigrams. These two bigrams contribute four different nodes (unigrams), namely (health, official, clinical, trials). It will 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

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 1 displays a wordcloud from the top-100 features of the training set. The figure shows proof of how relevant the features selected using the TFIDF algorithm.

Figure 1: shows a wordcloud representation of the top-100 features selected using TFIDF. Here we how observe relevant the features selected such as coronavirus, health, outbreak, covid, pandemic, social distancing, novel, spread, etc.

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. 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 prune 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 TFIDF features selected, and Figure 2 shows the training model after construction.

Number of Nodes	Number of Edges	Average Degree
412	471	2.2864

Table 2 describes the structural properties of the network training model: number of nodes, number of edges, and average node degree

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

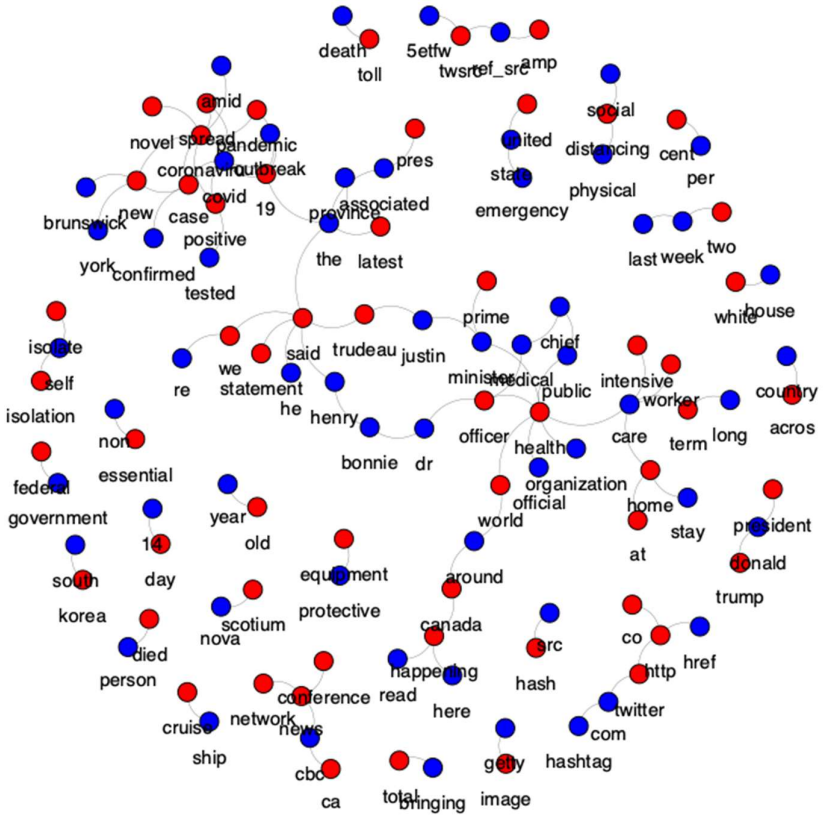

Figure 1 shows a simplified version of the network training model that is constructed from 100 features, which were originally extracted from the news articles. Here we see terms such as coronavirus outbreak, Covid 19, novel coronavirus, pandemic, positive cases

2.2. *The Design of NeoNet Classifier*

The previous step explained how a network-based training model is derived from a given set of news articles. Here we present the algorithmic steps that leads to labeling a new article that is yet to be seen by our algorithm. The algorithm is controlled by a configuration parameter which we called: minimum support which is inspired by the Apriori algorithm [43–45]. The minimum support guarantees a certain number of bigrams to be present in each article, otherwise, it will be labeled as suspicious. It ensures that the article contains sufficient contents that contributes to the training model. If the article does not meet this condition it will not be classified as SAFE. Clearly, an article that does not have a minimum number of TDIDF features also communicate significant facts worthy of reading [46]. 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 becomes significant in long vs short article. In long articles, it is expected to have a higher number of TFIDF features than shorter articles. If the minimum support parameter is set too low, this percentage helps correct this issue and ensure that news articles are not classified as "SAFE" if they should be classified as "SUSPICIOUS". The NeoNet algorithmic steps are described below and also is expressed in pseudocode in Algorithm 1.

1. *Set the minimum support parameter*

2. *For each new article to be classified*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

3. *Preprocess as previously described*
4. *Extract TFIDF 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: $mini\_conf$, the minimum confidence ratio r needed
Require: G, a graph training model
Initilize $min\_sup(d) = v$, for all $d \in \mathcal{D}$
Initilize $min\_conf(d) = r$, for all $d \in \mathcal{D}$
Initilize $positive(d) = 0$
Initilize $sup\_count(d) = 0$
Initilize $conf\_count(d) = 0$
repeat
  for each $d \in \mathcal{D}$ **do**
    $bigram\_list \leftarrow extract\_bigrams(d)$
    for each $bigram \in G\_list$ **do**
      if $left\_unigram \in G$ **then**
        add $G \leftarrow right\_bigram$
      else
        if $right\_unigram \in G$ **then**
          add $G \leftarrow right\_bigram$
        end if
      end if
      if $sup\_count \geq min\_sup$ & $conf\_count \geq min\_conf$ **then**
        $positive \leftarrow +1$
        $d \leftarrow COVID19$ label
      end if
    end for
  end for
until no more documents to classify

Algorithm 1 shows the steps of the NeoNet algorithm in pseudocode staring from when the bigram features are extracted until a classification label is 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 is to come up with a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

threshold that produces the best outcome. The minimum support parameter is 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 guides the algorithm to identify a reasonable threshold. A very high number of bigrams would lead to classifying articles that are extremely similar in content. As a result, the algorithm would miss articles that belong to the COVID-19 class, but less similar to the training model. On the other hand, a very low threshold would lead to classify any article with a slight overlap as positive and would make the algorithm not precise. We used the following min-sup configurations [5, 10, 15, 20, 25] bigrams. Figure 3 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 are matched against the training set and produced a precision value between (99.6%-100%). If it is 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 classification precision that fluctuates between (97.99%- 99.20%). Each fold is plotted using five different curves, one for each minimum support value.

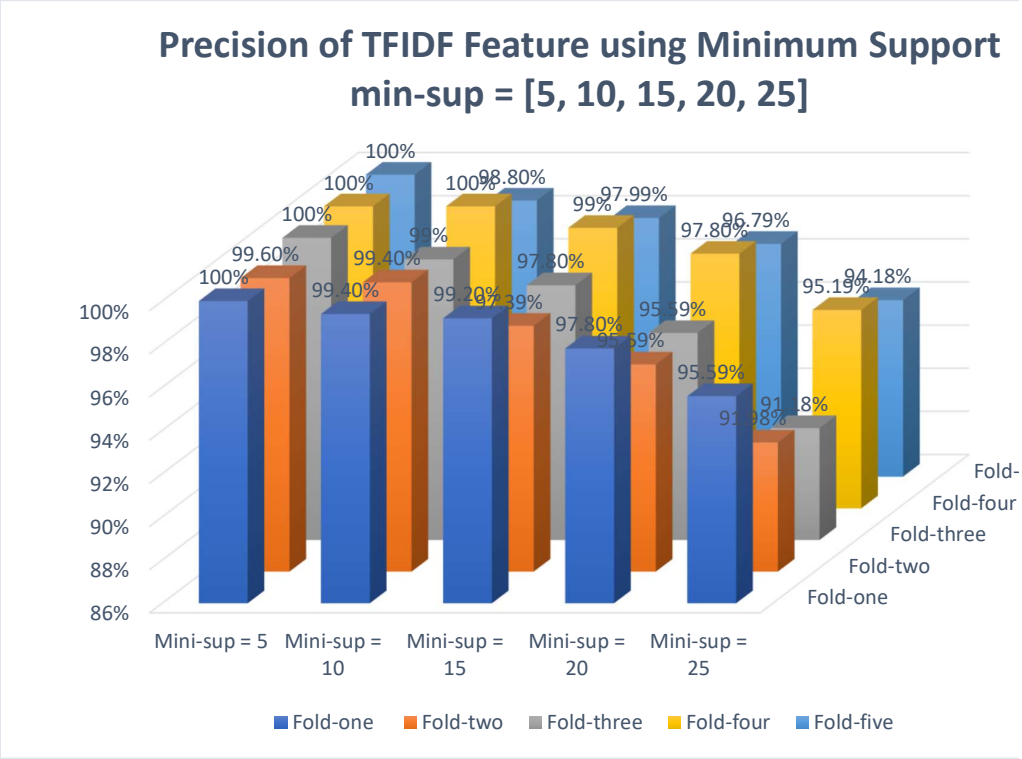

Figure 2 shows the performance of NeoNet with different configurations of the min-sup parameter. The figure shows 5 different configurations of the min-sup [5, 10, 15, 20, 25]. The plots are displayed from left to right respective to the values of the configurations

We have showed above how NeoNet can be controlled using the min support to make 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 don't necessarily utilize a similar notion of the min

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

support/confidence, we set both parameters to (min-sup=15). The algorithm is 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 the 5-folds. Figure 4 below shows how NeoNet's performance (shown on the far left of the x-axis) outperforms all other algorithms unless a perfect classification results is achieved by both algorithms (e.g., NeoNet vs Neural Net). The diagram shows a common theme: Except for NeoNet, Fold-3 (depicted in grey) appears to be scored the lowest among all the algorithms. Another noteworthy observation is that Fold-1 and Fold-5 also appear to be scored the highest (a 99.2%) among several algorithms which include Stochastic Gradient Descend, SVM, Random Forests, and Neural Nets. This is as close as it gets when their precision is compared with NeoNet. However, when all algorithms failed to achieve a perfect precision, NeoNet have shown dominance and outperformed all algorithms. We conducted the experiment using the Orange Data Mining Toolbox in Python46

A reasonable explanation for the outstanding performance that NeoNet has demonstrated is the feature selection part when training the algorithm. Clearly, the TFIDF bigram features are 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 are extremely indicative of the domain being analyzed. Additionally, the training model that is used is a natural extension of the individual bigram features because it naturally forms a network that accepts terms in common and rejects terms that don't contribute to the overall model. On the other hand, all prominent algorithms are driven from a model that relies on the individual terms and their frequencies. They all fail to integrate the associations with other terms. Clearly, such integration of these two characteristics (TFIDF bigram features and a network model) have indeed contributed to an exceptional performance demonstrated by NeoNet algorithm that is presented here.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2021 doi:10.20944/preprints202106.0482.v1

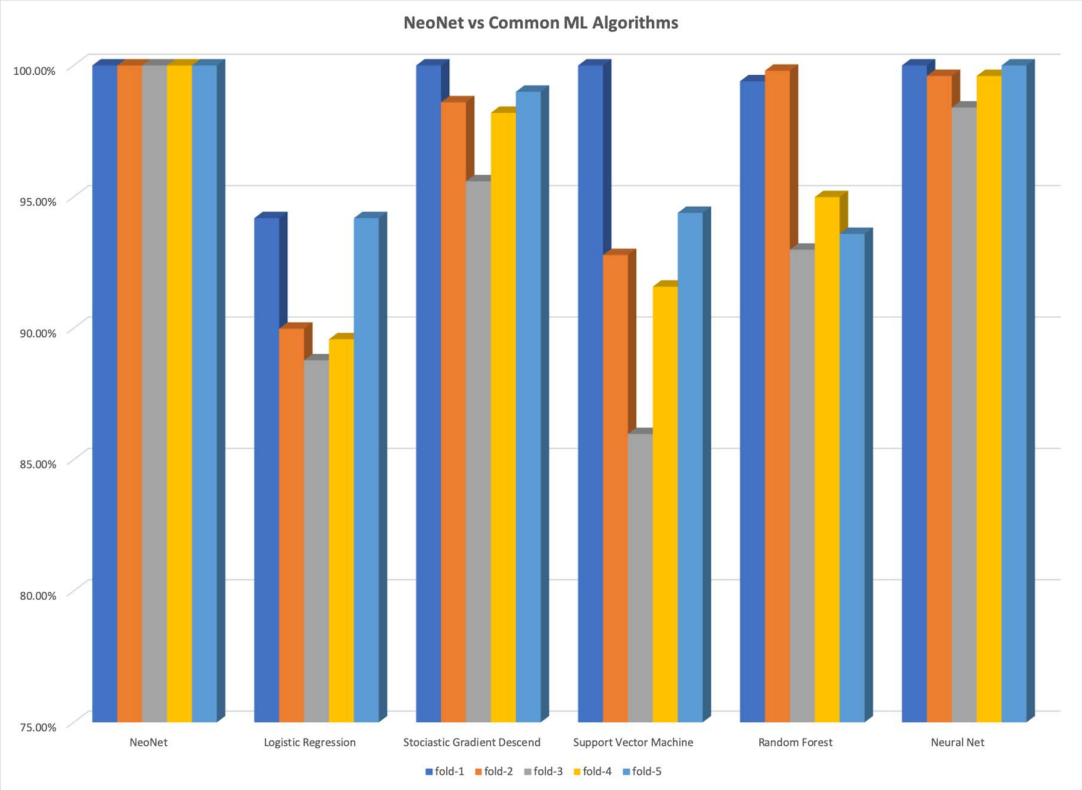

Figure 3 shows the performance analysis of NeoNet vs the most prominent Machine Learning algorithms. The algorithm performance is proving exceptional as it accurately classifies 5-folds dataset with a perfect precision surpassing Neural Net, SVM, Logistic Regression

4. Discussion and Future Direction

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 an evidence of a disinformation that existed in a publication, which eventually 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 also DailyMail news article [47] 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 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, encouraging knowledge refinement and quality improvement processes. Our research here has taken such recommendations into serious level 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

 doi:10.20944/preprints202106.0482.v1

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 have presented a novel machine learning algorithm that utilized knowledge refinement produced by natural language processing to produced 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 is published on Kaggle [36] (p10), 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, 25] and 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 it selects relevant and high-quality features that represents 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 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 that is based on TFIDF features for training its model, have taken the text classification from 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 will perfectly function on all other text sources, not only the news. Specifically, we expect the algorithm to function the exact same way, and without any modifications, to classify medical publications. The setting of the minimum support will be calibrated by experimentation. In this case, the algorithm may be trained on PubMed abstracts and doctor's notes to provide COVID-19 specific features that may only be clinical. Our reason to believe that it will be successful is that the algorithm is trained using bigram features. Such features are highly significant in the context of doctor's notes and publications since they may reference entities such as organs, disease, gene, protein, indication, symptom, etc. Eventually, the training model will be rich, and suspected sources will fail to classify positively against the model. To conclude, we produced an algorithmic approach that shows promise to fight the COVID-19 infodemic because of its powerful demonstration and flexibility in classifying news articles as "SAFE" or "DISPUTED". We also believe that training

the algorithm using scientific literature and doctor's notes may eliminate suspicious publications such as the one that promoted vitamin D. This is yet to be explored in future publications.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

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

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. You might choose to exclude this statement if the study did not report any data.

Acknowledgments: The authors tremendously thank Mrs. Regis O'Connor for providing her expertise producing some of the the graphics of this work. The authors also would like to thank Eszter Szenes and Zuzana Mikulecká for the valuable discussions.

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 JY, Muñoz-Perez MJ, Robledo-Vega C, Pérez-Zepeda MF, Soto-Vega E. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health Surveill*. 2020;6(2):e18444. doi:10.2196/18444

2. Hou Z, Du F, Zhou X, et al. Cross-Country Comparison of Public Awareness, Rumors, and Behavioral Responses to the COVID-19 Epidemic: Infodemiology Study. *J Med Internet Res*. 2020;22(8):e21143. doi:10.2196/21143

3. Hu Z, Yang Z, Li Q, Zhang A. The COVID-19 Infodemic: Infodemiology Study Analyzing Stigmatizing Search Terms. *J Med Internet Res*. 2020;22(11):e22639. doi:10.2196/22639

4. Moon H, Lee GH. Evaluation of Korean-Language COVID-19–Related Medical Information on YouTube: Cross-Sectional Infodemiology Study. *J Med Internet Res*. 2020;22(8):e20775. doi:10.2196/20775

5. Rovetta A, Bhagavathula AS. Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. *J Med Internet Res*. 2020;22(8):e20673. doi:10.2196/20673

6. Rovetta A, Bhagavathula AS. COVID-19-Related Web Search Behaviors and Infodemic Attitudes in Italy: Infodemiological Study. *JMIR Public Health Surveill*. 2020;6(2):e19374. doi:10.2196/19374

7. 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(5):1094-1099.

8. Tangcharoensathien V, Calleja N, Nguyen T, 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(6):e19659. doi:10.2196/19659

9. Gazendam A, Ekhtiari S, Wong E, et al. The "Infodemic" of Journal Publication Associated with the Novel Coronavirus Disease. *J Bone Joint Surg Am*. 2020;102(13). doi:10.2106/JBJS.20.00610

10. Okan O, Bollweg TM, Berens EM, 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(15). doi:10.3390/ijerph17155503

11. 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(8):e19311. doi:10.2196/19311

12. Dong W, Tao J, Xia X, et al. Public Emotions and Rumors Spread During the COVID-19 Epidemic in China: Web-Based Correlation Study. *J Med Internet Res*. 2020;22(11):e21933. doi:10.2196/21933

13. Stephens M. A geospatial infodemic: Mapping Twitter conspiracy theories of COVID-19. *Dialogues Hum Geogr*. 2020;10(2):276-281. doi:10.1177/2043820620935683

14. Islam MS, Sarkar T, Khan SH, et al. COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *Am J Trop Med Hyg*. 2020;103(4):1621-1629. doi:10.4269/ajtmh.20-0812

15. 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(5):327-328. doi:10.1097/MEJ.0000000000000713

16. Henrina J, Lim MA, Pranata R. COVID-19 and misinformation: how an infodemic fuelled the prominence of vitamin D. *Br J Nutr*. Published online undefined/ed:1-2. doi:10.1017/S0007114520002950

17. 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. doi:10.1016/j.ijinfomgt.2020.102201

18. Eysenbach G. How to Fight an Infodemic: The Four Pillars of Infodemic Management. *J Med Internet Res*. 2020;22(6):e21820. doi:10.2196/21820

19. 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(12):1285-1293. doi:10.1038/s41562-020-00994-6

20. Twitter to start removing COVID-19 vaccine misinformation. AP NEWS. Published December 16, 2020. Accessed December 26, 2020. https://apnews.com/article/misinformation-immunizations-coronavirus-pandemic-085cc1b49a5d488026f2e59d8f32d590

21. Braşoveanu AMP, Andonie R. Semantic Fake News Detection: A Machine Learning Perspective. In: Rojas I, Joya G, Catala A, eds. *Advances in Computational Intelligence*. Lecture Notes in Computer Science. Springer International Publishing; 2019:656-667. doi:10.1007/978-3-030-20521-8_54

22. Soon WM, Ng HT, Lim DCY. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Comput Linguist*. 2001;27(4):521-544. doi:10.1162/089120101753342653

23. Mackey TK, Li J, Purushothaman V, et al. 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(3):e20794. doi:10.2196/20794

24. Fei Liu, Feifan Liu, Yang Liu. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In: *2008 IEEE Spoken Language Technology Workshop*. ; 2008:181-184. doi:10.1109/SLT.2008.4777870

25. 4. Relationships Between Words: N-grams and Correlations - Text Mining with R [Book]. Accessed December 26, 2020. https://www.oreilly.com/library/view/text-mining-with/9781491981641/ch04.html

26. Qiang G. *An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification*. doi:10.1109/ICCRD.2010.160

27. Zhang H, Li D. Naïve Bayes Text Classifier. In: *2007 IEEE International Conference on Granular Computing (GRC 2007)*. ; 2007:708-708. doi:10.1109/GrC.2007.40

28. Meyer D, Leisch F, Hornik K. The support vector machine under test. *Neurocomputing*. 2003;55(1):169-186. doi:10.1016/S0925-2312(03)00431-4

29. Suthaharan S. Support Vector Machine. In: Suthaharan S, ed. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Integrated Series in Information Systems. Springer US; 2016:207-235. doi:10.1007/978-1-4899-7641-3_9

30. What is a support vector machine? | Nature Biotechnology. Accessed December 25, 2020. https://www.nature.com/articles/nbt1206-1565

31. Aphiwongsophon S, Chongstitvatana P. Detecting Fake News with Machine Learning Method. In: *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. ; 2018:528-531. doi:10.1109/ECTICon.2018.8620051

32. Ahmed H, Traore I, Saad S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I, Woungang I, Awad A, eds. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Lecture Notes in Computer Science. Springer International Publishing; 2017:127-138. doi:10.1007/978-3-319-69155-8_9

33. Conroy NK, Rubin VL, Chen Y. Automatic deception detection: Methods for finding fake news. *Proc Assoc Inf Sci Technol*. 2015;52(1):1-4. doi:https://doi.org/10.1002/pra2.2015.145052010082

34. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries. :4.

35. Ibrishimova MD, Li KF. A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing. In: Barolli L, Nishino H, Miwa H, eds. *Advances in Intelligent Networking and Collaborative Systems*. Advances in Intelligent Systems and Computing. Springer International Publishing; 2020:223-234. doi:10.1007/978-3-030-29035-1_22

36. COVID-19 Open Research Dataset Challenge (CORD-19). Accessed December 25, 2020. https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

37. pandas - Python Data Analysis Library. Accessed December 25, 2020. https://pandas.pydata.org/

38. TextBlob - Google Search. Accessed July 9, 2020. https://www.google.com/search?q=TextBlob&oq=TextBlob&aqs=chrome..69i57j35i39j69i59j0l5.2340j0j4&sourceid=chrome&ie=UTF-8

39. Hamed AA, Ayer AA, Clark EM, Irons EA, Taylor GT, Zia A. Measuring climate change on Twitter using Google's algorithm: perception and events. *Int J Web Inf Syst*. 2015;11(4):527-544. doi:10.1108/IJWIS-08-2015-0025

40. Dey A, Pal A. Prim's algorithm for solving minimum spanning tree problem in fuzzy environment. *Ann Fuzzy Math Inform*. 2016;12:419-430.

41. Wang W, Huang Y, Guo S. Design and Implementation of GPU-Based Prim's Algorithm. *Int J Mod Educ Comput Sci*. 2011;3(4):55-62. doi:10.5815/ijmecs.2011.04.08

42. "Clear-cut" evidence coronavirus drug remdesivir works, Fauci says. NBC News. Accessed December 25, 2020. https://www.nbcnews.com/health/health-news/coronavirus-drug-remdesivir-shows-promise-large-trial-n1195171

43. Toivonen H. Apriori Algorithm. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning*. Springer US; 2010:39-40. doi:10.1007/978-0-387-30164-8_27

44. Al-Maolegi M, Arkok B. An Improved Apriori Algorithm for Association Rules. *ArXiv14033948 Cs*. Published online March 16, 2014. Accessed October 24, 2020. http://arxiv.org/abs/1403.3948

45. Perego R, Orlando S, Palmerini P. Enhancing the Apriori Algorithm for Frequent Set Counting. In: Kambayashi Y, Winiwarter W, Arikawa M, eds. *Data Warehousing and Knowledge Discovery*. Lecture Notes in Computer Science. Springer; 2001:71-82. doi:10.1007/3-540-44801-2_8

46. 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)*. Association for Computational Linguistics; 2019:110-114. doi:10.18653/v1/D19-6617

47. "Alarmingly high" proportion of British people are vitamin D deficient | Daily Mail Online. Accessed December 25, 2020. https://www.dailymail.co.uk/sciencetech/article-9068299/Alarmingly-high-proportion-British-people-vitamin-D-deficient.html