

Classifying COVID-19 Disinformation on Twitter using a Convolutional Neural Network

Mohamad Nabeel and Christine Große^{1a}

Department of Information Systems and Technology, Mid Sweden University, Holmgatan 10, Sundsvall, Sweden

Keywords: Deep Learning, COVID-19, Twitter Data, Intelligent Systems, Disinformation, Fake News, Convolutional Neural Network, CNN.


Abstract: Disinformation regarding COVID-19 is spreading rapidly on social media platforms and can cause undesirable consequences for people who rely on such content. To combat disinformation, several platform providers have implemented intelligent systems to detect disinformation and provide measurements that apprise users of the quality of information being disseminated on social media platforms. For this purpose, intelligent systems employing deep learning approaches are often applied, hence, their effectivity requires closer analysis. The study begins with a thorough literature review regarding the concept of disinformation and its classification. This paper models and evaluates a disinformation detector that uses a convolutional neural network to classify samples of social media content. The evaluation of the proposed deep learning model showed that it performed well overall in discriminating the fake-labelled tweets from the real-labelled tweets; the model yielded an accuracy score of 97.2%, a precision score of 95.7% and a recall score of 99.8%. Consequently, the paper contributes an effective disinformation detector, which can be used as a tool to combat the substantial volume of disinformation scattered throughout social media platforms. A more standardised feature extraction for disinformation cases should be the subject of subsequent research.

1 INTRODUCTION

The concept of disinformation is nothing new. Deceptive advertising, government propaganda, deepfakes and forged documents are a few of the many methods that can be employed in business and politics to achieve the objectives of disinforming an audience (Fallis, 2015). Disinformation, which gained particular attention during the Second World War, has been employed by foreign powers to mislead people and disrupt businesses using information warfare. Although disinformation has been and still is a common means utilised by foreign states to project (political) power onto other states, the development of information technologies in recent years has accelerated this issue. Therefore, spreading false information has become a more prevalent tool for disseminating inaccurate and misleading information for political purposes (Fallis, 2015).

Most recently, along with the COVID-19 pandemic, an 'infodemic' has emerged in which a variety of information regarding COVID-19 has been

published, composed of both accurate and inaccurate information (Song et al., 2021). This infodemic has influenced the public to mistrust official information and to employ treatments that have endangered people's health (Song et al., 2021). Numerous rumours, constituting a risk to people's health and the political stability of states, have spread rapidly through a variety of social media platforms. It is therefore important to identify whether the information being distributed regarding COVID-19 is true or not to warn media users whether the content they are reviewing is suspicious. Such detection of disinformation contributes to efforts to reduce the number of false claims that are spread through the Internet, which, in turn, may lead to a reduction in undesirable consequences. At present, information systems are attempting to address this issue. In particular, providers are employing intelligent systems that can classify disinformation, also known as *fake news*, on social media platforms. One example is the social media platform Twitter, which anyone can access to post tweets, read and like tweets

^a <https://orcid.org/0000-0003-4869-5094>

as well as retweet content on the platform. In the case of Twitter, intelligent systems using machine learning and deep learning models analyse content and likes to combat fake news. The models can incorporate the likes in a hybrid approach using two neural networks (Kumar et al., 2019; Umer et al., 2020) and machine learning algorithms such as support vector machine and naïve Bayes (Reddy et al., 2019). Although models that classify disinformation do exist, most of them focus solely on the text and title content (e.g. Qawasmeh et al., 2019; Rath and Basak, 2020; Verma et al., 2019); only a few consider other features (Sahoo and Gupta, 2021).

Consequently, this study aims to fill this gap by contributing a disinformation detector for social media content that is able to classify whether a piece of information is true or false. For this purpose, we implement a deep learning approach that applies a convolutional neural network (CNN) to improve the detector's ability to classify information from a given data set. The focus is also on extracting detailed features that can be fed into the model. Hence, a thorough literature review precedes the construction of the model to examine the characteristics and concepts of disinformation that provide the solid foundation for the CNN.

2 THEORETICAL FRAMEWORK

2.1 Disinformation

Three important features characterise disinformation according to Fallis (2015). First, disinformation is considered to be a type of information. The exact definition of disinformation depends on which analysis of information is adopted (ibid.). There are many analysis approaches; however, the central feature of information is that it “represents some part of the world as being a certain way” (ibid). In particular, information is an artefact that has semantic or representational content. Other research has recognised objects or documents that contain certain descriptions or summaries as such forms of representation (Buckland, 1991). However, more features are necessary to distinguish disinformation from information. Second, disinformation is a type of misleading information (Fallis, 2015), which means that disinformation is likely to lead to or create false beliefs. It should be noted that disinformation does not necessarily have to mislead someone to be classified as disinformation; its intended purpose will still be regarded as disinformation regardless of whether or not the receiver believes in the message.

However, disinformation always puts people at risk. Third, disinformation is misleading by intention (Fallis, 2015). This feature is what clearly distinguishes disinformation from misinformation, since the latter covers content that is considered to consist of honest mistakes or overly subtle satire. Hence, disinformation is regarded as misleading information that has the purpose of misleading.

In addition, a systematic literature review recently studied the phenomenon of disinformation (Kapantai et al., 2021). The review examined existing typologies of false information, particularly the underlying motives, facticity and verifiability of disinformation. Table 1 presents disinformation types and their suggested underlying motives.

Table 1: The unified typology framework for disinformation (Kapantai et al., 2021).

Dimensions/ Measurement	Motive		
	Profit	Ideological	Psychological
Clickbait	X		X
Conspiracy theories		X	X
Fabrication	X		
Misleading connection	X		
Hoax	X		
Biased or one-sided	X		
Imposter	X		
Pseudoscience	X		X
Rumours	X		
Fake reviews	X		
Trolling	X		

The *motives* for employing disinformation include financial, ideological and psychological intentions. The *facticity* of disinformation is suggested to be mostly true, mostly false or false (Kapantai et al., 2021). The label ‘mostly true’ means that the statement or parts of the statement are accurate and contain facts that require additional clarification or information. The label ‘mostly false’ means that the statement or parts of it are inaccurate; it contains true elements but ignores critical facts that could give the receiver a different impression if provided. The label ‘false’ means that the whole statement is inaccurate. Finally, the binary dimension *verifiability* clarifies whether the disinformation is verifiable or not. The authors of the review argued that the facticity of all the disinformation types is mostly true, that they are verifiable, and that most of them are designed for profit purposes (Kapantai et al., 2021). Despite the fact that these findings leave room for further questions to be posed, this paper follows the

assumption that the suggested values for motive and facticity can be viewed similarly for the COVID-19 disinformation under study.

To differentiate between information, misinformation and disinformation, several attributes of information quality can be employed, wherein the following criteria are of particular value (e.g. Große, 2021; Tudjman and Mikelic, 2003).

- *Authority* – extent to which the author(s) and sponsor(s) as well as copyrights are disclosed.
- *Accuracy* – extent to which information is correct, flawless and certified free of error.
- *Objectivity* – extent to which information is unbiased, unprejudiced and impartial.
- *Timeliness/Currency* – extent to which information, source and context are up to date and updateable by direct communication.
- *Completeness* – extent to which information is of sufficient breadth, depth and scope.
- *Representation* – extent to which information is well-organised, concise and consistent as well as interpretable, readable and considerate of the human ability to analyse information.

In particular, disinformation should be considered in the following cases: (a) the original author(s) and source(s) remain hidden, (b) the information is not verifiable through evidence or facts, (c) the information reflects a personal point of view, (d) the information is out-of-date or without options for updating discussions, (e) the information is improperly restricted, and (f) the information is inconsistent and confusing (cf. Tudjman and Mikelic, 2003).

Due to the increased spread of disinformation on social media platforms, researchers have noted the necessity for a framework that identifies anomalous or suspicious digital information even without the knowledge of anomalous samples. A practical guideline has proposed that the best option for detecting fake news is to focus on the news sources, such as a popular web page or an unknown domain revealed by suspicious tokens in the URL (Zhang and Ghorbani, 2020). Advisory sections on web sites, such as ‘about us’ or ‘disclaimer’ sections, could be used as a credibility indicator (ibid). To assess the truthfulness of the content, a user could check the supporting resources that a particular author provides, the date of the news and its recurrence in other feeds. In addition to practical recommendations, deep learning algorithms have become increasingly common in fake news detection on social media platforms in recent years. Such approaches, based on data-mining techniques, not only rely on handcrafted textual features but can also capture the hidden

implications of the contextual and author information over time. Computation units and extra hidden layers are suggestions for further improvement of this method (Zhang and Ghorbani, 2020).

Zhang and Ghorbani (2020) suggest three main types of features for fake news detection:

- A creator/user-based feature set,
- A news content-based feature set,
- A social context-based feature set.

The first and most significant set includes user-profiling features, such as the verification, description and data registration of the user; user-credibility features, such as the number of posts that connects to the users; and user-behaviour features. The second set of features constitutes a powerful tool for fake news analysis, which includes the news topic, the number of special tags or symbols in the entire message and external links. This set includes linguistic/syntactic-, style- and visual-based features. The third set includes network-based features that analyse a user’s educational background, habits, location and sports, for example. In addition, distribution-based features help to capture distinct diffusion patterns in the news, which include the number of retweets or reposts of the original post. Temporal-based features complete the third set; they analyse how frequently a user posts news and at what time or on which day of the week.

This paper limits its scope to features from the first and second set, namely creator/user-based features and news content-based features.

2.2 Deep Learning

In recent years, deep learning algorithms have progressed and achieved success in speech and visual object recognition. In particular, such approaches have shown promising results in the context of fake news (Goldani et al., 2021; Kaliyar et al., 2020; Zhang and Ghorbani, 2020). Unlike conventional machine learning techniques, which require handcrafted feature extractions, deep learning algorithms can process raw data and automatically discover representations. In general, deep learning seeks to imitate natural learning mechanisms by creating an artificial neural network (ANN). While extensive knowledge concerning deep learning is available, this paper limits this section to the essential background that is relevant for the development of the disinformation classifiers.

Figure 1 illustrates the basic structure of an ANN, which consists of an input layer of neurons, hidden layers of neurons, and an output layer of neurons.

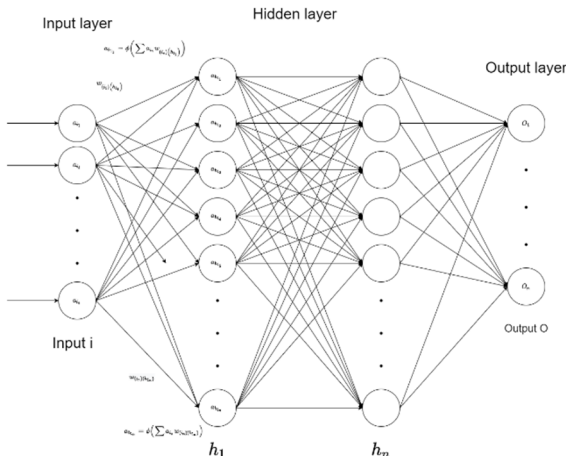


Figure 1: The architecture of an ANN.

Each input neuron holds an input value, extracted from the feature values of a sample in the data set, and can produce a single output with a weight value added for each of the neurons in the next layer of neurons (Grossi and Buscema, 2007).

Assuming that the neurons are fully connected, the values that the neurons of the next layer will inherit depend on the sum of the output values from each of the predecessor neurons. They are then used as input for an activation function φ noted as

$$a_{l_n} = \varphi \left(\sum_{j=1}^n a_{k_j} w_{k_j l_n} \right) \quad (1)$$

where a_{l_n} is the value of the neuron n within the layer l , a_{k_j} is the value of the neuron j within the predecessor layer k and $w_{k_j l_n}$ is the weight connection from neuron j within the predecessor layer k to the neuron n within the layer l

The activation function can vary; logistic functions or Gaussian functions are common. A bias value b can also be added at the end of the sum notation if increasing the likelihood of a neuron being active or inactive is desired. A deep neural network that employs multiple layers between the input and output layers is a subset of an ANN.

2.3 Convolutional Neural Network

A CNN is a subset of a deep neural network, typically used for image classification and recognition as well as natural language processing (NLP). In comparison to the main approach employed in the context of handwriting and speech recognition cases – recurrent neural networks (Zhang and Ghorbani, 2020) – CNNs are computationally cheap. Unlike ANNs, CNNs use

convolutional layers in a weight-sharing scheme, which improves its learning efficiency (Mujeeb et al., 2019).

Given the extracted feature values in an NLP case, the initial input of these values (e.g. words) is converted into numerical values that a computer can comprehend. Then, each of the values is converted into n -dimensional vectors, which together are further concatenated into a matrix, where l numbers of filters w with a fixed kernel size k are applied (Gu et al., 2018). In general, CNN architectures are stacked with three main types of layers: the convolution layer, pooling layer and fully connected layer (e.g. Stanford University, 2021).

For NLP cases, the one-dimensional convolutional layer is the most appropriate. This architecture means that the size k is the length of the filter – a vector that contains k elements including different numerical values. The filter performs a dot product operation on each row of the matrix that is aligned with the filter. Different filters, which contain different numerical values, apply the dot operation on different rows of the matrix. Once the convolutions are complete, the matrices are pooled down and flattened using the pooling layer, which can then be sent to the feedforward network (Gu et al., 2018). A common pooling layer for NLP is called *global max pooling* which employs the maximum value of each of the vector outputs. Figure 2 depicts the typical CNN architecture.

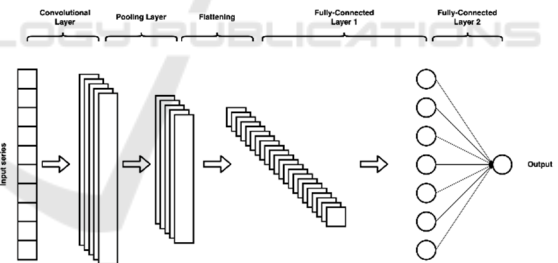


Figure 2: The architecture of a CNN (Lewinson, 2020).

2.4 Related Work

Research investigating the use of deep learning for fake news detection has increased substantially over recent years. For example, Ruchansky et al. (2017) proposed a model that combines the text of a news article, its source, the users promoting it and the response it receives to classify fake news. The model yielded a scoring accuracy of 0.892 and an F-score of 0.894 in the analysis of a Twitter data set. As another example, Wang et al. (2018) developed an end-to-end framework, called the Event Adversarial Neural Network, which can derive event-invariant features

and, thus, aids in fake news detection on newly emerging events. Together with an event discriminator, a multi-modal feature extractor and a fake news detector, this network can learn transferable features for unseen events to detect fake news.

Although these and other studies perform deep learning and provide the reasonable scores that are expected from successfully implemented deep learning approaches, these studies are also rather generic. They apply a general perspective on fake news classification and do not investigate a specific subject. In addition, these studies fail to focus on the concept and characteristics of disinformation. This consideration raises questions about whether the selections made during model development are appropriate and how well the approaches perform in the context of digital disinformation.

Consequently, this study anchors its model development in the elaboration of disinformation, as previously detailed. This proceeding assists with finding more relevant features that can be used to detect fake news along with the text content. In particular, the model is slanted toward the specific news topic of COVID-19.

3 METHODOLOGY

This study applied a four-step methodology inspired by the CRISP-DM model (see Chapman et al., 2000).

- Data collection
- Data pre-processing
- Implementation of the model
- Performance evaluation

The following tools were used during the study. Python constituted the programming language, the scikit-learn library was employed for evaluations and the Keras library was used in the implementation of the neural network model. The Pandas library was applied for the management and pre-processing of the data set.

3.1 Data Collection

The study utilised data sets from Kaggle and Twitter. The Kaggle² site is an online community of machine learning practitioners and data scientists and provides a collection of data sets. This study extracted a data set comprised of a collection of COVID-19 fake news,³ published as short messages (tweets) on Twitter.⁴

² <https://www.kaggle.com/>

³ <https://www.kaggle.com/arashnic/covid19-fake-news>

To extract the necessary data from the data set, a developer account was created on Twitter. Initially, the tweets from the data sets were dehydrated, meaning that they only contained the unique tweet IDs. Once extracted, the tweets were assembled and opened. The prepared data set was comprised of 12,469 samples; approximately half of them were labelled as 'fake' (numerically labelled as 1) and the other half were labelled as 'real' (numerically labelled as 0).

3.2 Data Pre-processing

After assembling the dataset, we selected the features to include in the disinformation detector from those available in the Twitter API. The feature selection was based on the previously outlined fundamentals of disinformation. Upon further consideration, a time difference feature was added to the user-based feature set. This specific feature concerns the difference between the date of the tweet and the date when the user account was created and was measured in hours. This additional feature functioned as a bot indicator and considered the reliability of the account. In particular, the following features were included:

Creator/user-based Features

- the place of the user
- the URLs posted in descriptions to gain additional information about the user
- the source URLs posted in tweets which can be considered as the source of information
- the verification of the user
- the time difference

News Content-based Features

- the tweet text of the user

3.3 Implementation of the Model

The implementation of the model was comprised of two steps: tokenisation and the training of the model. At the outset, the text must be tokenised to suit the neural network classifier. To this end, the text was demarcated. Each word was labelled with a certain index value, the more frequently the word appeared, the lower the value it received. The tokenised text sample was separated from the other feature values and fed into an embedding layer. The embedding layer placed words that may have similar meanings in the same category so that the neural network could interpret them as equal.

⁴ <https://help.twitter.com/en/new-user-faq>

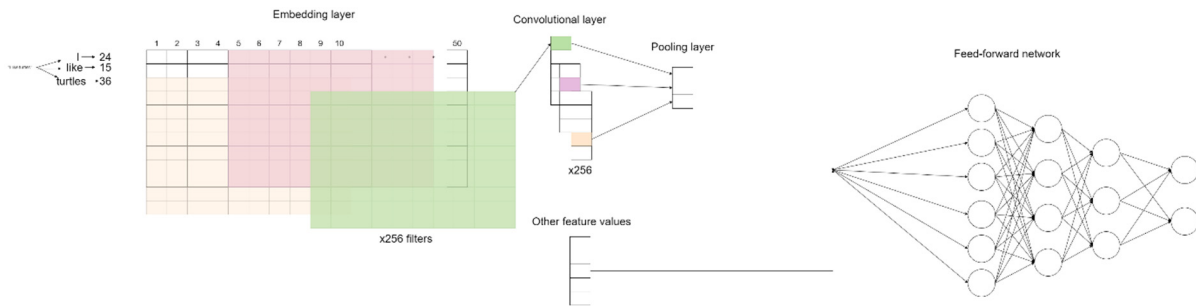


Figure 3: The schematic classification model.

Each word was converted into a 50-dimensional vector. The rectified linear units activation function was applied for the two dense layers and Sigmoid for the output layer (see e.g. Agarap, 2019; Han and Moraga, 1995).

As Figure 3 illustrates, the implementation used one convolution layer with 256 filters after the embedded text and before a global max-pooling layer. Once fully connected, the text layer was concatenated with the other feature values before the dense layers finally classified the tweets and provided the final output. The ratio split was set to the optimal train-test ratio, which is inversely proportional to the square root of the number of free adjustable parameters p (Amari et al., 1997; Guyon, 1997), in accordance with Equation (2).

$$\text{validation (\%)} = \frac{1}{\sqrt{p}} * 100 \quad (2)$$

For the training of the model, p was set to 6, which gave a train-test split ratio of 59-41. Hence, approximately 59% of the samples were utilised to train the disinformation detector and 41% to test it.

3.4 Performance Evaluation

First, the confusion matrix in Table 3 supports the evaluation of the detector performance regarding the correct classification of each sample from the dataset.

Table 2: Confusion Matrix.

Prediction	Fake	Real
Fake	True Positive (TP)	False Positive (FP)
Real	False Negative (FN)	True Negative (TN)

The predicted samples that the detector truly labels as *fake* are counted by TP, FP counts the predicted samples wrongly labelled as *fake* which are *real*, FN counts the predicted samples wrongly labelled as *real* which are *fake*, and TN counts the predicted samples truly labelled as *real*. The evaluation applies the following scores to assess the detector's performance:

accuracy, precision, recall and F1 [see Equation (3) – (6)]. In addition, the evaluation from various threshold settings utilises a receiver operating characteristics (ROC) curve. The ROC curve is a graphical display of the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis (Kumar and Indrayan, 2011) [see Equation (7) – (8)].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

$$\text{TPR} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (8)$$

The values of TPR and FPR can range from 0 to 1. The focus is on the area under the curve (AUC), which is an effective way to assess the validity of the test (Kumar and Indrayan, 2011). The value of AUC can range from 0 to 1. Whereas a value of 1 indicates a perfectly accurate test, a zero value means that the test classified all the samples incorrectly. Finally, the evaluation applies a 10-fold cross-validation on the detector model to prove the consistency of the scores.

4 RESULTS

Table 3 summarises the results of the performance evaluation.

Table 3: The performance of the disinformation detector.

	Accuracy	Precision	Recall	F1
Score	97.7%	95.7%	99.8%	97.7%

The proposed CNN-based disinformation detector achieved an accuracy score of 97.2%. This result means the model labels approximately 98 out of 100 samples correctly and two incorrectly. The achieved precision score is 95.7%, which indicates that out of all the samples labelled as fake by the detector, 95.7% were correctly labelled as fake. The achieved recall score of 99.8% demonstrates that out of all the samples that were actually fake, 99.8% of them were correctly labelled as fake by the detector. The F1 score – the harmonic mean of precision and recall – of 97.7% signifies that the model is not biased so far.

In addition, Table 4 displays the confusion matrix of the predictions made by the proposed CNN model on the validation set. Considering that the model performs on a binary classification dataset which has a balanced distribution of labels, the predictions of TP and TN also appear balanced.

Table 4: The performance of the disinformation detector.

Prediction	Fake	Real
Fake	1,826 (TP)	49 (FP)
Real	263 (FN)	1,603 (TN)

Training the disinformation detector model and evaluating it with the 10-fold cross-validation, as Table 5 demonstrates, yielded an average accuracy score of 98.1%, a precision score of 98.9%, a recall score of 99.5% and an F1 score of 99.2%. The standard deviation showed no indication of inconsistencies regarding the scores that the model achieved.

Table 5: The 10-fold cross-validation scores.

K-folds	Accuracy	Precision	Recall	F1
Fold 1	99.2%	98.9%	99.5%	99.2%
Fold 2	98.0%	96.1%	99.9%	98.0%
Fold 3	99.2%	98.8%	99.7%	99.2%
Fold 4	97.8%	95.7%	99.6%	97.8%
Fold 5	96.3%	99.6%	93.0%	96.2%
Fold 6	99.0%	98.5%	99.6%	99.1%
Fold 7	97.9%	95.9%	99.9%	97.9%
Fold 8	99.2%	98.5%	99.8%	99.2%
Fold 9	97.5%	95.3%	99.9%	97.6%
Fold 10	96.6%	99.5%	93.5%	96.4%
Mean	98.1%	97.7%	98.5%	98.1%
Standard deviation	1.03%	1.62%	2.64%	1.08%

Figure 4 displays the ROC curve, which provides an overview of the results of the thresholds. The mean

AUC of 0.98 confirms that the developed disinformation detector model is performing well overall in discriminating the *fake*-labelled tweets from the *real*-labelled tweets.

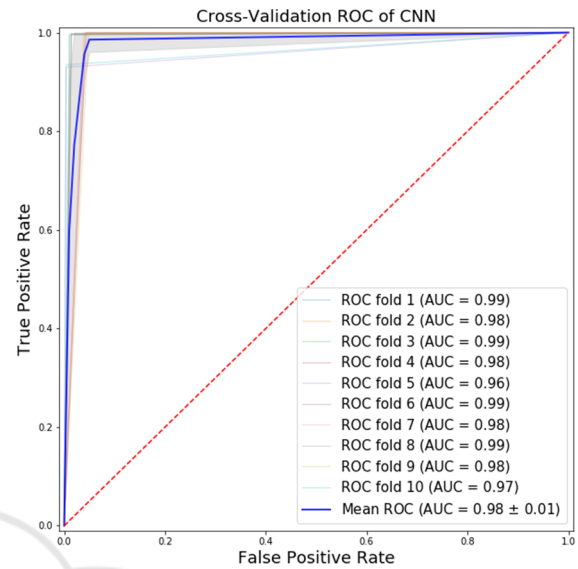


Figure 4: The ROC curve and the AUC values.

5 DISCUSSION

The aim of this study was to provide a disinformation detector for social media content using a deep learning architecture that considers parameters based on the theoretical concept and characteristics of disinformation. Although existing deep learning models offer a complex architecture to combat the overwhelming disinformation feeds in social media, many models have difficulties with applying their full capabilities and being evaluated accordingly, due to imbalanced label distribution and insufficient samples in data sets. This study thus confirms the accuracies of current CNN models for fake-news detection, such as those recently demonstrated by Kaliyar et al. (2020) and Goldani et al. (2021).

The research for this study further identified a lack of feature standardisation in the existing data sets with regard to disinformation because different APIs provide different features extracted from the related database. However, the findings of this study accurately contribute to efforts to define appropriate features that should be included in a disinformation detector for social media content (see e.g. Sahoo and Gupta, 2021). To this end, the concrete impact that each of the features has on the model’s accuracy requires further assessment.

Although deep learning models can provide high scores and promising results, it is difficult to evaluate exactly how the model picks, chooses and prioritises values in its architecture. This uncertainty raises the question of whether the model emulates reality in an appropriate manner. In addition, the initial selection of features that the model includes should be subject to future investigations which include the role of model developers and providers of deep learning services in fake news detection.

Since a CNN facilitates both the processing of more data in the network and quicker conclusions than other network approaches, utilising a CNN for the disinformation detector model provided a fast and efficient prediction with a high-level result. However, despite comprehensive evaluations, the study has not scrutinised in detail to what extent the high performance scores are related to the thorough feature selection. Future research could thus include an evaluation of the effect that each of the features has on the accuracy score.

In addition, an analysis of the relationship between the promising results and the composed architecture requires further comparison with a different model such as a recurrent neural network. In comparison to earlier research, which has focused on fewer features in a more complex neural network (e.g. Qawasmeh et al., 2019; Rath and Basak, 2020; Verma et al., 2019), the disinformation detector presented here demonstrates that even a simpler neural network combined with more features can achieve high levels of accuracy. This finding indicates that multi-feature extraction and feature engineering are promising approaches for fake news detection, which confirms the results of recent research (Sahoo and Gupta, 2021).

Despite a dedicated focus on COVID-19 news, the proposed approach is still limited to a rather generalised approach towards disinformation. Other neural networks, such as Event Adversarial Neural Networks (Wang et al., 2018) could be utilised to delve into special cases of disinformation campaigns.

This study proposed the disinformation detector with a broad audience in mind. One group of interest might be plug-in programme developers who wish to implement a fake news detector to raise the awareness of social media users regarding COVID-19. Another group of users that this study envisions may be analysts who work with competitive intelligence to understand the current state of COVID-19 and the latest risks and opportunities that have occurred on the Internet.

6 CONCLUSIONS

This paper models and evaluates a disinformation detector that uses a CNN network to classify samples of social media content, especially Twitter messages. The intention was to contribute a reliable approach to automatically classifying a piece of digital information as true or false. Therefore, the study began with a thorough literature review with regard to the concept of disinformation and its classification. Based on the literature review, we designed and implemented a disinformation detector, which yielded promising evaluation results. Nonetheless, although the model exhibited impressive scores, further assessment is advisable. Considering both the rapidly growing number of posts and tweets on social media platforms and their variety, a larger data set with more samples is needed to facilitate a detailed analysis of the detector's capabilities. Despite many deep learning models classifying accurately and, in some cases, also efficiently, the black box aspect of these models impedes a nuanced interpretation of their operation. Future research could address the development of standardised feature extraction for disinformation cases, which could facilitate the extraction of data sets that reflect reality regarding domain-specific cases. Another suggestion for future research efforts targets the elaboration of disinformation classifiers. For example, analyses could evaluate to what extent models developed for specific subjects, such as the COVID-19 news in this paper, affect classification performance.

REFERENCES

- Agarap, A. F. (2019). Deep Learning using Rectified Linear Units (ReLU). *arXiv, 1803.08375v2*. Retrieved 2021-09-01, from <https://arxiv.org/pdf/1803.08375.pdf>.
- Amari, S., Murata, N., Muller, K.-R., Finke, M., and Yang, H. H. (1997). Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks, 8*(5), 985–996.
- Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science, 42*(5), 351–360.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., et al. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS. Retrieved 2021-08-31, from <https://www.the-modeling-agency.com/crisp-dm.pdf>.
- Fallis, D. (2015). What Is Disinformation? *Library Trends, 63*(3), 401–426.
- Goldani M.H., Safabakhsh, R., and Momtazi, S. (2021). Convolutional neural network with margin loss for fake

- news detection. *Information Processing & Management*, 58(1).
- Große, C. (2021). Enhanced Information Management in Inter-organisational Planning for Critical Infrastructure Protection: Case and Framework. In *Proceedings of the 7th International Conference on Information Systems Security and Privacy*. SCITEPRESS - Science and Technology Publications, 319–330.
- Grossi, E., and Buscema, M. (2007). Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, 19(12), 1046–1054.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77(11), 354–377.
- Guyon, I. (1997). *A Scaling Law for the Validation-Set Training-Set Size Ratio*. AT&T Bell Laboratories, Retrieved 2021-09-01, from <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1337>.
- Han, J., and Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In J. Mira and F. Sandoval (Eds.), *Lecture Notes in Computer Science: Vol. 930. From natural to artificial neural computation. IWANN 1995*. Berlin: Springer.
- Kaliyar, R. K., Goswami, A., Narang, P., and Sinha, S. (2020). FNDNet – A deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61, 32–44.
- Kapantai, E., Christopoulou, A., Berberidis, C., and Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5), 1301–1326.
- Kumar, R., and Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian pediatrics*, 48(4), 277–287.
- Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., and Akbar, M. (2019). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2).
- Lewinson, E. (2020). *Python for Finance Cookbook: Over 50 Recipes for Applying Modern Python Libraries to Financial Data Analysis*: Packt.
- Mujeeb, S., Alghamdi, T. A., Ullah, S., Fatima, A., Javaid, N., et al. (2019). Exploiting Deep Learning for Wind Power Forecasting Based on Big Data Analytics. *Applied Sciences*, 9(20), 4417.
- Qawasmeh, E., Tawalbeh, M., and Abdullah, M. (2019). Automatic Identification of Fake News Using Deep Learning. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 383–388.
- Rath, P. K., and Basak, R. (2020). Automatic Detection of Fake News Using Textual Entailment Recognition. In *2020 IEEE 17th India Council International Conference (INDICON)*. IEEE, 1–6.
- Reddy, P. S., Roy, D., Manoj, P., Keerthana, M., and Tijare, P. (2019). A Study on Fake News Detection Using Naïve Bayes, SVM, Neural Networks and LSTM. *Journal of Advanced Research in Dynamical & Control Systems*, 11(06), 942–947.
- Ruchansky, N., Seo, S., and Liu, Y. (2017). CSI - A Hybrid Deep Model for Fake News Detection. In E.-P. Lim, M. Winslett, M. Sanderson, A. Fu, J. Sun, S. Culpepper, et al. (Eds.), *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 797–806.
- Sahoo, S. R., and Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100(3), 106983.
- Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., et al. (2021). Classification aware neural topic model for COVID-19 disinformation categorisation. *PLoS one*, 16(2), e0247086.
- Stanford University (2021). *CS231n Convolutional Neural Networks for Visual Recognition*. Retrieved 2021-08-31, 2021, from <https://cs231n.github.io/convolutional-networks/>.
- Tudjman, M., and Mikelic, N. (2003). Information Science: Science about Information Misinformation and Disinformation. In: *InSITE Conference*. Informing Science Institute.
- Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., et al. (2020). Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM). *IEEE Access*, 8, 156695–156706.
- Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., et al. (2018). EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *KDD2018. August 19-23, 2017, London, United Kingdom*. New York, NY: Association for Computing Machinery Inc. (ACM).
- Verma, A., Mittal, V., and Dawn, S. (2019). FIND: Fake Information and News Detections using Deep Learning. In *2019 12th International Conference on Contemporary Computing (IC3)*, 1–7.
- Zhang, X., and Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102025.