






Article

An Explainable Fake News Detector Based on Named Entity Recognition and Stance Classification Applied to COVID-19

Giorgio De Magistris ¹, Samuele Russo ², Paolo Roma ³, Janusz T. Starczewski ⁴ and Christian Napoli ^{1,*}

- ¹ Department of Computer, Automation and Management Engineering, Sapienza University of Rome, via Ariosto 25, 00185 Roma, Italy; demagistris@diag.uniroma1.it
- ² Department of Psychology, Sapienza University of Rome, via dei Marsi 78, 00185 Roma, Italy; samuele.russo@uniroma1.it
- ³ Department of Human Neurosciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Roma, Italy; paolo.roma@uniroma1.it
- ⁴ Department of Intelligent Computer Systems, Czestochowa University of Technology, al. Armii Krajowej 36, 42-200C Czestochowa, Poland; janusz.starczewski@pcz.pl
- * Correspondence: cnapoli@diag.uniroma1.it

Abstract: Over the last few years, the phenomenon of fake news has become an important issue, especially during the worldwide COVID-19 pandemic, and also a serious risk for the public health. Due to the huge amount of information that is produced by the social media such as Facebook and Twitter it is becoming difficult to check the produced contents manually. This study proposes an automatic fake news detection system that supports or disproves the dubious claims while returning a set of documents from verified sources. The system is composed of multiple modules and it makes use of different techniques from machine learning, deep learning and natural language processing. Such techniques are used for the selection of relevant documents, to find among those, the ones that are similar to the tested claim and their stances. The proposed system will be used to check medical news and, in particular, the trustworthiness of posts related to the COVID-19 pandemic, vaccine and cure.

Keywords: natural language processing; named entity recognition; CNN; fake news; COVID-19; vaccines; explainable artificial intelligence



Citation: De Magistris, G.; Russo, S.; Roma, P.; Starczewski, J.T.; Napoli, C. An Explainable Fake News Detector Based on Named Entity Recognition and Stance Classification Applied to COVID-19. *Information* **2022**, *13*, 137. <https://doi.org/10.3390/info13030137>

Academic Editor: Kostas Vergidis

Received: 6 February 2022

Accepted: 2 March 2022

Published: 7 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Nowadays social media has reshaped the mass communication ecosystem. Individual news creators with no confirmed reputation can reach wide audiences on news networks because of the absence of the verification of data, such as third-party filtering [1]. These are the ideal conditions for the proliferation of fake news. Fake news can be defined as the deliberate presentation of false and misleading claims as real facts [2]. They may cause serious impact and even damage to society since they may be intentionally forged to manipulate the orientation of people regarding important themes. The recent COVID-19 pandemic has exposed the fragility of the modern mass media, also highlighting the importance of the awareness regarding the illness and the related therapy and prevention [3]. During the still ongoing worldwide COVID-19 pandemic, therefore, fake news have also become a serious risk for the public health, especially when false or misleading information is spread regarding the nature of this illness and its cure. This study introduces an automatic detection system that finds a set of documents from verified sources that support or disprove a claim. This work is motivated by the fact that, while much effort has been devoted to the development of automated systems, according to our knowledge, still few investigations have been dedicated to the explanation of the obtained results. We believe that explanations are needed in order to convince people about the mendacity of a claim. For this reason, we developed a system that leaves the final judgment to the user, presenting them all the evidence, rather than being a straightforward binary classifier. This approach

follows the recent thread of the Explainable AI (XAI), which aims to make AI systems more understandable by humans by providing explanations [4]. We used deep language models in order to encode the documents into a feature vector space, then a deep convolutional neural network is employed to classify the stances of two documents.

The rest of the paper is structured as follows: in Section 2 there is an overview of other existing approaches. Section 3 describes the different datasets used in the project. The proposed method is described in details in Section 4. The final results are reported in Section 5 and discussed in Section 6, in which also further improvements are introduced. In Section 7 conclusions are drawn.

2. Related Works

Thanks to the advances in natural language processing (NLP) several automated fake news detection systems have been developed in recent years. The task is generally formulated as a supervised classification problem [5]. Many approaches are based on machine learning models for classification such as support vector machine (SVM), naive Bayes classifier (NBC), logistic regression (LR) and random forest classifier (RFC). Further, neural networks have been widely used, especially recurrent neural networks (RNN) and convolutional neural networks (CNN), often used jointly as in [6], and networks based on attention mechanisms [7,8]. Another important aspect is the data representation. Common choices used to represent an entire sequence as a vector are bag of words and tf-idf embeddings while the most used word level distributed representations are word2vec [9] and GloVe [10].

Fake news detection systems can be classified into two macro categories: news content models and social context models [11]. The first category includes: (1) the style-based methods, which use language statistics to find “deception cues” using only the textual content of the document [12]; (2) knowledge-based methods, in which the claim is compared with a knowledge base of trusted content. Social context models exploits linked data that are typical of social networks content. This kind of information can be either used alone or to enrich the information contained in the text of the claim. This category can be further subdivided into: (1) stance-based methods, in which the other users viewpoint (stance) is used to assess the veracity of the claim. For example [13,14] completely disregard the textual information and use only the stances. The first uses the sequence of stances of the post comments to learn a hidden Markov model, while the second uses logistic regression on the binary vector of stances (like or dislike) from a fixed set of users. (2) Propagation-based methods consider the network formed by the posts and employ graph optimization techniques [15–17]. There are also successful techniques that mix multiple approaches. For example dFEND [18], detects fake news in social networks using both the content of the message and the comments from other users. It is based on a sentence–comment co-attention subnetwork that, in addition to the binary classification, reports the top-k check-worthy sentences and the top-k comments that contributed to the result.

Our approach consists of searching evidence for fake news detection using the results of the stance classification, which is the task of assessing what side of the debate an author is on from text [5]; however, we only use a knowledge base of trusted sources with no network information. Similarly to our approach, the authors of [19] introduced a fake news detection system that compares the claim with a set of headers of news articles using stance classification; however, the information contained in the header could not be sufficient to prove or disprove a claim. Another aspect that should be considered is how to efficiently search for the relevant articles. The proposed method considers these aspects in order to improve the fake news classification methods based on pure textual information and stance classification.

3. Dataset

The proposed fake news detection system is based on the comparison between the text to be verified (from now on it is called “query”) and some trustful material. To reduce

the number of documents that contribute to the classification, the query is first assigned to a macro-category and then to a sub-category (a subset of the macro-category). The macro-category and sub-category datasets are used for these classification tasks. The BBC news dataset for text classification [20] was used for the macro-categories entertainment, sport, tech, business and politics while the medicine macro-category was added manually using scientific articles collected from PubMed and PMC. Further, the medicine sub-category dataset was built manually, starting from a collection of articles from PubMed and PMC and categorized into the following eight sub-categories: COVID-19, Cancer, Bone Disease, Depression, Nutrition, Hepatitis, Sexually Transmitted Diseases and Heart Disease. In a first attempt, we used the same sub-category dataset to validate the query with trusted material; however, scientific papers differed too much both in the structure, the content and the language from the query. Secondly, in present time, the news articles related to COVID-19 are more likely containing facts that can be easily distorted or misreported since they are more related to subjective factors such as disbelief and gossips without any scientific background rather than “bone disease” or “heart disease”, which are more related to objective factors (aging, genetics, level of healthcare, etc). For these reasons, we decided to use a coronavirus news dataset as the trustworthy reference for the implementation of the fake news detection system, narrowing the focus to the specific topic of COVID-19. This dataset (aggregated, analyzed and enriched by AYLIEN using AYLIEN’s News Intelligence Platform) contains 1,673,353 news articles collected during the period from November 2019 to July 2020 from about 440 global verified sources. Each article has many annotations, but for our purpose we used only the title, the body and the URL (such that the interested user can read its full content). To train the stance classification network, we employed the FNC-1 dataset (stance dataset), used in the Fake News Challenge competition [21]. It contains labeled entries where the first element is the headline of a news, the second element is the body and the label is one among the following: (1) Discuss, if the two elements are related but without making claims about the topic, (2) Agree, if the two elements are related and make statements about the topic that are in agreement and (3) Disagree, if the two elements are related and make disagreement statements about the topic.

4. Method

The proposed fake news detection system is structured into multiple modules, the workflow is represented in Figure 1. The user submits a query that is classified into a macro-category. Based on the result of this classification, the system selects only the relevant articles for the specific topic (or decide to discard the query if it does not match any macro-category). In this implementation, only queries about medicine will be processed, but the same approach can be generalized to any other macro-category, provided the existence of a collection of trusted documents on that topic. Generally a macro-category can be partitioned into many subcategories. A second classification step assigns to the query one of the sub-categories to further reduce the number of documents that will contribute to the evaluation. As a further phase, named entity recognition (NER) is used to filter, both from the documents and the query, the sentences that do not contain named entities (more details are given in Section 4.2). In the next step, the query is compared with the relevant documents using document embedding and cosine similarity and the top-k most similar will be the candidate for the stance classification. The stance classification model (StanceNN) classifies the top-k documents with respect to the query with one of the labels: Agree, Disagree, Discuss or Unrelated.

In particular let $q \in R^d$ be the embedding of the query, and

$$X = \{x_i \in R^d \mid \text{cosine_similarity}(q, x_i) \geq t\} \quad (1)$$

the set of documents that are similar to q , where t is a fixed threshold. Then, the candidates for the stance classification are:

$$\hat{X} = \{x_i \in X \mid i \in [1, k] \wedge \text{cosine_similarity}(q, x_i) \geq \text{cosine_similarity}(q, x_{i+1})\} \quad (2)$$

The stance classification network learns the mapping from the embedding space to the labels space:

$$x_i \in \hat{X} \longrightarrow c_i \in \{Agree, Disagree, Discuss, Unrelated\} \tag{3}$$

The result of the stance classification allows us to partition the set \hat{X} of the k most similar documents into the sets:

$$\begin{aligned} \hat{X}_{agree} &= \{x_i \in \hat{X} \mid StanceNN(x_i) = Agree\} \\ \hat{X}_{disagree} &= \{x_i \in \hat{X} \mid StanceNN(x_i) = Disagree\} \\ \hat{X}_{discuss} &= \{x_i \in \hat{X} \mid StanceNN(x_i) = Discuss\} \\ \hat{X}_{unrelated} &= \{x_i \in \hat{X} \mid StanceNN(x_i) = Unrelated\} \end{aligned} \tag{4}$$

Finally the output of the fake news classifier is:

$$Assessment(q) = \begin{cases} True & \text{if } |\hat{X}_{agree}| > |\hat{X}_{disagree}| \\ Fake & \text{if } |\hat{X}_{disagree}| > |\hat{X}_{agree}| \\ Unknown & \text{Otherwise} \end{cases} \tag{5}$$

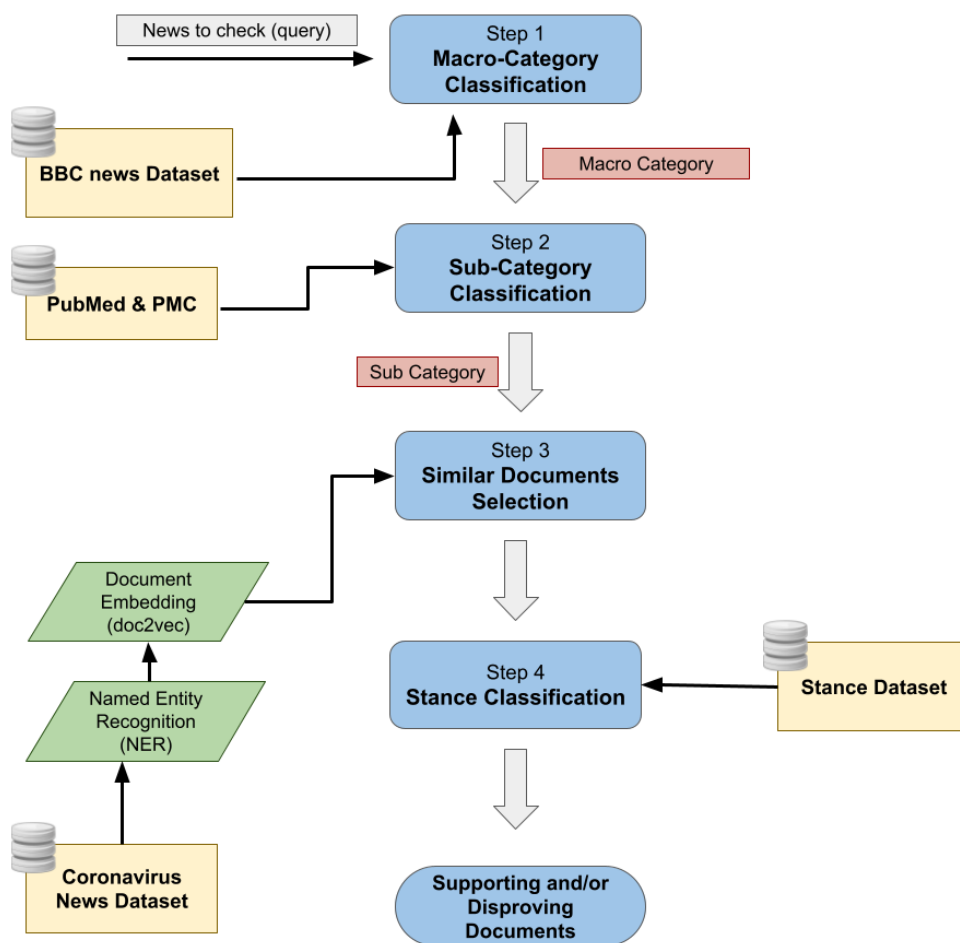


Figure 1. Fake news detection system overview.

4.1. Query Classification

The two classification problems introduced in Section 4 differ only in the dataset used for training, in particular the macro-category and sub-category datasets described in Section 3 are used, respectively, for the macro-category and sub-category classification. Many machine learning algorithms have been tested and, among those, support vector machine, K nearest neighbor, random forest and naive Bayes gave all excellent results. The precision for each of those models are reported in Table 1 both for the macro-category and sub-category classification.

In order to achieve the best performance, an ensemble method with a majority-voting scheme [22] was introduced. In order to train these models the documents are transformed into vectors using the tf-idf representation. Each vector has one entry for each term in the vocabulary that is defined as:

$$TF - IDF_{d,t} = TF_{d,t} \log \frac{N}{1 + DF_t} \quad (6)$$

where $TF_{d,t}$ is the frequency of term t into document d , DF_t is the number of documents with the term t and N is the number of documents. The vocabulary must be the same for the documents and the query hence the terms that appear in the query but not in the vocabulary used for training are ignored.

Table 1. Precision for the macro-category (top) and sub-category (bottom) classification on the validation set.

	SVM	KNN	RF	NB
Macro-category classification Precision				
Business	97%	96%	94%	97%
Entertainment	99%	99%	98%	99%
Medicine	99%	99%	97%	100%
Politics	99%	90%	96%	95%
Sports	99%	98%	97%	99%
Tech	99%	93%	97%	98%
Sub-category classification Precision				
Bones Diseases	100%	93%	100%	100%
Cancer	88%	99%	92%	95%
COVID-19	81%	99%	67%	65%
Depression	100%	81%	95%	100%
Heart Diseases	100%	96%	100%	96%
Hepatitis	93%	89%	97%	90%
Nutrition Diseases	74%	73%	93%	82%
STDs	100%	93%	100%	87%

4.2. Named Entity Recognition

Named entity recognition is the task of identifying and classifying names in text. NER is used to solve many NLP tasks such as text understanding, question answering, summarization, information retrieval, machine translation, knowledge base construction, etc. Different approaches have been used in the literature to solve this problem, they can be categorized into three main classes [23]: (1) rule-based approaches where hand crafted rules based on language statistics are used to label the parts of the text; (2) unsupervised learning approaches, where the words are assigned to clusters based on their context; (3) feature-based supervised learning approaches, where some models learn discerning rules from discriminative features, such as SVM, decision trees and maximum entropy-based models, while others learn the probability distribution that better fits the training data, such as hidden Markov models or conditional random fields; (4) deep-learning-based approaches, which are the ones adopted in this study, because they have the advantage to automatically

learn complex features directly from data and it is often possible to use models that are pre-trained on huge corpus and then fine tune them for the specific task.

A general deep learning model for NER is composed by a context sensitive encoder that represents each word as a vector depending on its surroundings and a decoder that assigns the label to each vector. Possible choices for the encoder are recurrent neural networks, bidirectional RNNs, convolutional neural networks and transformers, while the decoder can be a standard multilayer perceptron with Softmax activation for multi-class classification, a conditional random field or a recurrent neural network. In this work, we used bidirectional encoder representations for transformers (BERT) [24] as encoder and a multilayer perceptron with Softmax activation as decoder. We adhered to the implementation provided by [25], because the model was first trained on a general corpus (English Wikipedia and BooksCorpus) and then fine-tuned on domain specific corpus (collected from PubMed and PMC). BERT is a bidirectional language model (the embedding of a word depends both on its left and right context) that is based on the transformer architecture. The transformer, introduced in [26] is a context-encoder network based solely on attention mechanisms, without employing neither recurrence nor convolution. It reached state-of-the-art results when fine-tuned for NER [23] and BioBERT (BERT fine-tuned on medical corpus) further improves the results of BERT for NER on medical texts [25].

In an early stage of development we implemented a very simple preprocessing for the stance dataset, consisting of tokenization only, converting to lower case and stripping of punctuation; however, such results are affected by the presence of non-existing words. For this reason we introduced a more advanced semantic filtering based on NER, in particular we split the document into sentences and filtered those that did not contain at least one named entity. This technique introduced many advantages: on the one hand the size of the dataset was reduced and on the other hand the embeddings of the documents were not affected by unimportant information.

4.3. Similar Documents Selection

After that the query is assigned to a sub-category, the next step consists in finding the top-k documents in the same sub-category that are more similar to the query. For this task doc2vec [27] and cosine similarity [28] are used. Doc2vec is a distributed document embedding that gave excellent results in the task of finding similar documents when combined with cosine similarity as shown in [29]. For this specific task we implemented and trained from scratch the distributed memory model of paragraph vectors, according to which the paragraph vector and word vectors are concatenated to predict the next word in a context. The model was trained on the *coronavirus news dataset* 100 epochs. At test time the word embedding matrix and the parameters of the model are fixed, while the paragraph vector is updated through gradient descent to provide a distributed representation of the document.

4.4. Stance Classification

The last step is the stance classification, that consists of finding the position of the top-k similar documents with respect to the query. Two architectures were implemented and tested on the stance dataset described in Section 3. The two architectures share the same word embedding that was obtained training doc2vec on the *coronavirus news dataset* as described in the previous section (Figure 2 shows some 3D representation of the embedding of some significant words obtained with PCA).

Each network has two input branches, one for the query and one for the document. The outputs of the two branches are then concatenated and further processed. The final layer has 3 neurons with Softmax activation function in order to output a probability distribution on the four labels: Unrelated, Agree, Disagree and Discuss. The first architecture is a bidirectional long short-term memory (LSTM) inspired by [30]. The two branches are composed by a single bidirectional LSTM layer with 128 hidden units, then the concatenation of the output vectors is passed through a single dense layer with 3 units. The

second architecture is a convolutional neural network inspired by [31,32]. Each branch is composed of three 1D convolutional layers, with 128 filters interleaved by max pooling layers with stride 1 to enhance the interesting feature extracted by the convolutional layers and to simplify the representation. The kernel size defines the number of consecutive words among which the network is able to discover dependencies, analogously to an n-gram model. Experimental evidences showed that a kernel with size 3 gave the best results (the details of the architecture are illustrated in Figure 3) The bidirectional LSTM achieved a 65% accuracy after 50 epochs of training while the CNN a 78% accuracy in about 700 epochs; however, for the bidirectional LSTM a single epoch took about 45 min for completion while for the CNN it took just few seconds. Anyway, it should not be excluded that with more training time the bidirectional LSTM could perform even better.

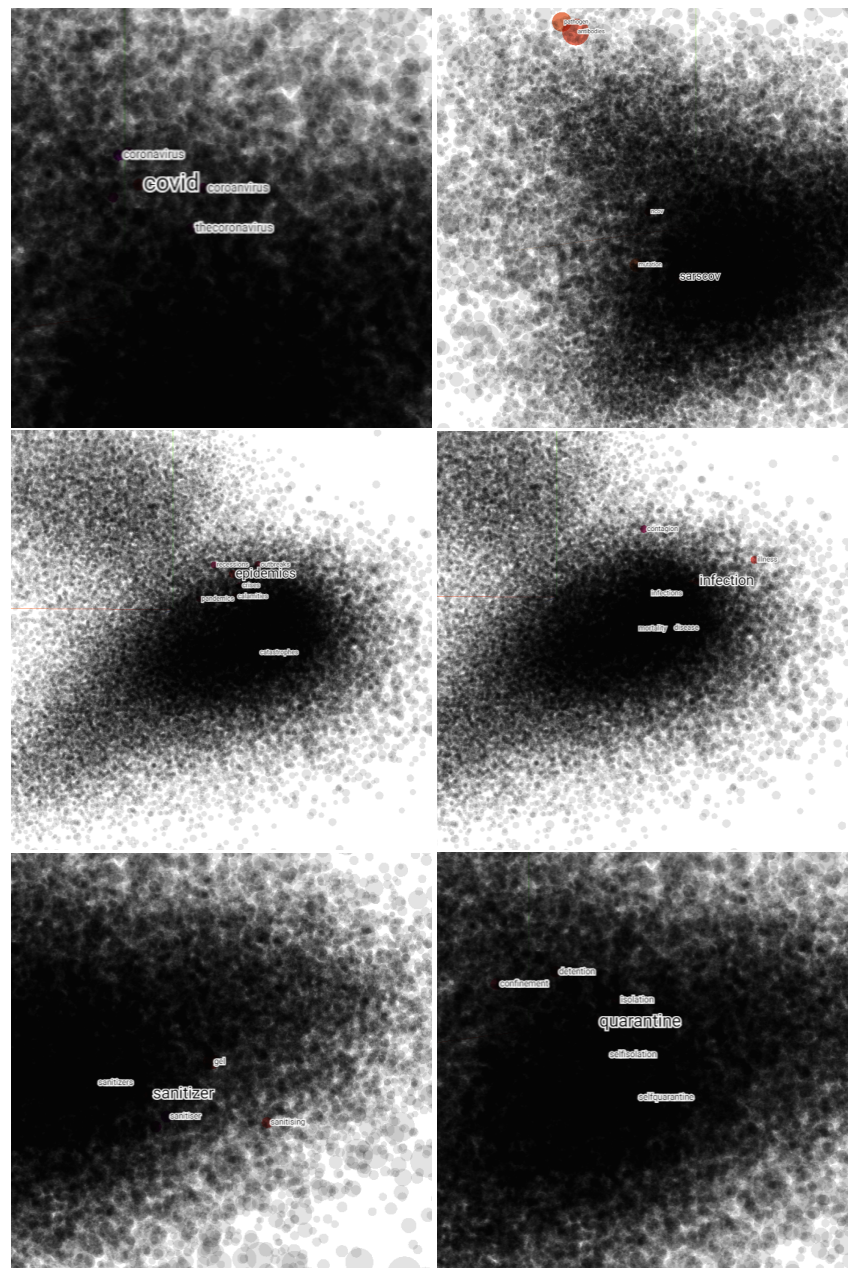


Figure 2. Word embeddings, obtained with doc2vec, projected onto a 3D subspace with principal component analysis (PCA). In each figure there is a word that is connected to the coronavirus and its closest neighbors, where the proximity measure is cosine similarity.

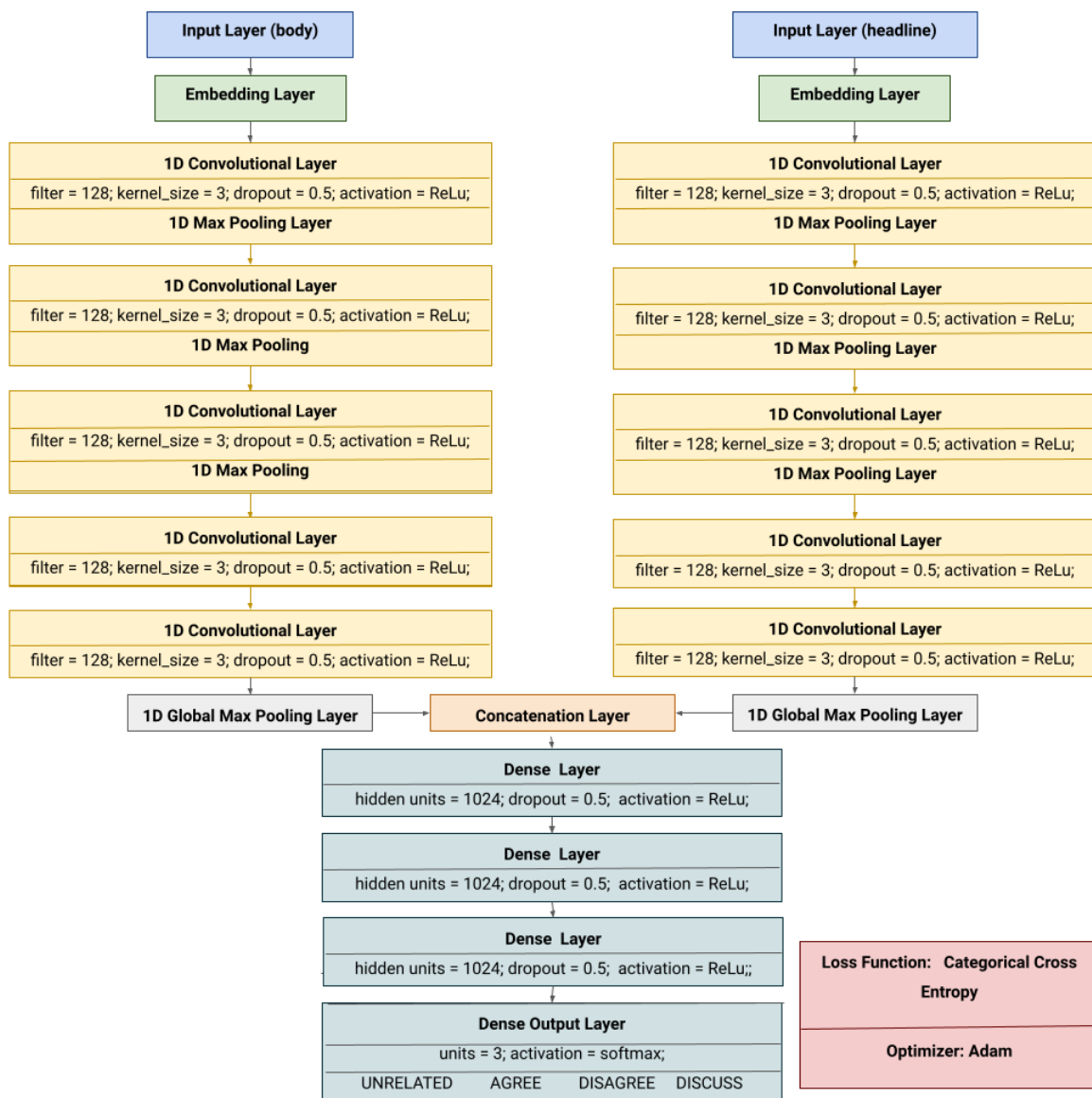


Figure 3. In the figure the architecture of the convolutional network for stance classification. The network has two inputs, one for the headline (the claim to be verified) and the other for the body (the similar document). The embedding layer is taken from the trained doc2vec model. The two branches have the same number of convolutional layers interleaved by pooling layer and the last layer is global max pooling, that takes the maximum value over the time dimension. The two branches are then concatenated and passed through three dense layers with reLu activation and finally the output layer with three neurons and Softmax activation function for multi-class classification.

5. Results

In this section we discuss the integration of the modules described in the previous section and the performances of the system. The precision for the two classification tasks (macro-category and sub-category classification) have already been shown in Section 4.1. The similar document selection (after that non relevant sentences have been filtered with NER) achieved good results (see Table 2 for some examples).

To validate the proposed method we fixed a number *k* of similar documents (we tried different values), and counted the numbers of documents that “agreed” and “disagreed” with the query. Then, if the first was greater than the second, the query was considered “authentic”, otherwise it was considered “fake”. For this purpose we used the CoVID19-FNIR [33] dataset, that contains labeled true and fake news about coronavirus. The results

however were not satisfactory and so we tried different solutions. By inspecting the results of the stance classification, it emerged that the majority of documents were labeled as *Unrelated*. This was probably due to the inability of the stance classification network (that obtained good results on the FNC-1 validation set) to generalize on a completely different dataset (in this case the *coronavirus news dataset* and the *CoVID19-FNIR* dataset). To improve the performances of the stance classification network we fine-tuned the model by manually labeling some similar documents. In particular, for each true or fake document from the CoVID19-FNIR dataset we used doc2vec to retrieve the 10 most similar documents from the *coronavirus news dataset*, and we provided the pairs of similar documents with the correct label signal to update the weights of the stance network. After the finetuning, the number of documents classified with the other labels increased considerably, however the number of documents labeled as *Agree* or *Disagree* was still too small to build a reliable classifier using just this information. Moreover, to reduce the effect of the *Unrelated* documents, we introduced a variable k . An initial value is set to 3, then, if 1 of the 3 documents is classified as *Unrelated*, then k is incremented by 1 until there are three similar documents that are all classified with the labels (*Agree*, *Disagree*); however, the idea of building the binary classifier counting the labels was discarded for two reasons: it was not reliable, since precision and recall were too low, and it did not give additional information to the user. Indeed, by inspecting the results of the stance classification it emerged that most of the time the system assigned meaningful labels, but often the real stance of the document with respect to the query is subtle. For this reason in the final solution the stance classification is used mainly to filter the similar documents that are unrelated with the query and to provide a rough categorization of the similar documents. Some examples are illustrated in Tables 3–5.

Table 2. Example of similar document selection. The query is taken from the *CoVID19-FNIR* dataset while the similar documents are the ones with the highest cosine similarity among all the documents in the *coronavirus news dataset*.

Query	
the indian embassy in tokyo has said that one more indian crew member on Diamond Princess has tested positive for covid	
Similar Documents	
Similarity	Text
0.64	3rd indian tests positive for coronavirus on ship off coast Japan the indian embassy in Tokyo has said that one more indian crew member on Diamond Princess, the ship stranded off coast Japan, has tested positive for COVID-19. The person has been hospitalized for treatment [. . .]
0.60	Four Indian crew members, who underwent tests for the coronavirus along with others still on board the cruise ship moored off the Japan coast, have tested positive for COVID-19, taking the total number of Indians infected with the virus on the vessel to 12, the Indian embassy said on Sunday. Passengers showing no signs of the deadly disease started disembarking the ship, Diamond Princess, after the quarantine period ended last week.
0.58	One more indian has been tested positive for the novel coronavirus disease. Covid on the quarantined cruise ship diamond princess off the japanese coast. The Indian embassy in Japan confirmed on friday this is the third confirmed case of indian nationals testing positive for the deadly coronavirus which has claimed lives of over [. . .]

Table 3. Examples of answers provided by the system for different input queries (case study n. 1).

Example 1				
Query (ground = fake)				
coronavirus has been found in broiler chicken				
Similarity	Stance	Explanation Text	url	
0.54	Disagree	The Poultry Farm Association on Saturday organized this Chicken Mela to dispel rumours that the bird is a carrier of Coronavirus. Vineet Singh, president of the Poultry Farm Association said that people had stopped eating chicken since the past one month, due to fear of Coronavirus. "We organized this Mela where we invited people to eat chicken. We wanted to tell them that Coronavirus is not caused by eating chicken, mutton or fish. We cooked over a thousand kilograms of chicken for the Mela and the entire stock was finished", he said. The Chicken Mela, held in front of the Gorakhpur railway station, proved to be a major crowd puller and left all roads leading to the railway station blocked for hours.	https://www.news18.com/news/india/free-chicken-mela-to-dispel-coronavirus-rumours-in-ups-gorakhpur-2520867.html (accessed on 4 January 2022)	
0.54	Disagree	A full plate of chicken dishes for Rs 30 may sound unbelievable but it was a dream come true for chicken lovers in Gorakhpur. The Poultry Farm Association on Saturday organized this Chicken Mela to dispel rumours that the bird is a carrier of Coronavirus. Vineet Singh, president of the Poultry Farm Association said that people had stopped eating chicken since the past one month, due to fear of Coronavirus. "We organized this Mela where we invited people to eat chicken. We wanted to tell them that Coronavirus is not caused by eating chicken, mutton or fish. We cooked over a thousand kilograms of chicken for the Mela and the entire stock was finished", he said.	https://www.hindustantimes.com/its-viral/gorakhpur-organizes-free-chicken-mela-to-shatter-coronavirus-myths/story-rtsaf0G5GvSmgopCjjLC9K.html (accessed on 4 January 2022)	
0.50	Discuss	Coronavirus is reported to have started from Wuhan, China and in the wet markets where people come every day to shop for meats. In these markets, people sell and buy all kinds of meats—chicken, seafood, mutton, sheep, pig and even snakes. Because of this very reason, people in India are doubting if they should eat seafood. To put a rest to this confusion, it has been said that its safe to eat seafood in India as no such link between sea animals and coronavirus has been established.	https://timesofindia.indiatimes.com/life-style/food-news/5-foods-linked-to-novel-coronavirus-and-the-truth/photostory/73935050.cms (accessed on 4 January 2022)	

Table 4. Examples of answers provided by the system for different input queries (case study n. 2).

Example 2				
Query (ground = true)				
an elderly chinese tourist hospitalised in france has died of the coronavirus covid				
Similarity	Stance	Explanation Text	url	
0.74	Agree	France's health minister on Saturday announced the first coronavirus death in Europe, an 80-year-old Chinese tourist who other French authorities say was initially turned away from two French hospitals when he first fell ill. Minister Agnes Buzyn said she was informed Friday night of the death of the patient, who had been in intensive care at Bichat Hospital in Paris after testing positive in late January. His daughter also tested positive for the virus that has spread across central China and was hospitalized. However, the health minister said she was doing well and should be leavin	https://www.france24.com/en/20200215-france-announces-first-coronavirus-death-outside-asia (accessed on 4 January 2022)	

Table 4. Cont.

Example 2		
0.66	Agree	<p>An 80-year-old Chinese tourist has died of the fast-spreading coronavirus in France, becoming the first fatality in Europe, French Health Minister Agnes Buzyn said on Saturday. France has recorded 12 cases of the virus, out of a global total of 67,000. The vast majority of those suffering from the virus are in China. The epidemic has killed more than 1500 people. Buzyn said she was informed on Friday that the patient, who had been treated at the Bichat hospital in northern Paris since 25 January, died of a lung infection due to the coronavirus. “This is the first fatality by the coronavirus outside Asia, the first death in Europe”, Buzyn told reporters. “We have to get our health system ready to face a possible pandemic propagation of the virus, and therefore the spreading of the virus across France”, she added.</p> <p>https://www.reuters.com/article/us-china-health-france-idUSKBN2090B0 (accessed on 4 January 2022)</p>
0.62	Agree	<p>France confirms fourth case of coronavirus in elderly Chinese tourist. France on Tuesday reported that a fourth person was infected with the coronavirus, an elderly Chinese tourist. Health Ministry director Jerome Salomon said the patient, hospitalized in Paris, was a Chinese tourist believed to be about 80 years old. “His medical situation is serious, as he is requiring resuscitation”, Salomon told reporters.</p> <p>https://www.reuters.com/article/us-china-health-france/france-confirms-fourth-case-of-coronavirus-in-elderly-chinese-tourist-idINKBN1ZR2CM?edition-redirect=in (accessed on 4 January 2022)</p>

Table 5. Examples of answers provided by the system for different input queries (case study n. 3).

Example 3		
Query (ground = fake)		
the asterix comic books and the simpsons predicted the coronavirus outbreak		
Similarity	Stance	Explanation Text
0.52	Agree	<p>The Simpsons fans are convinced Tom Hanks’ coronavirus diagnosis was predicted by his cameo in the 2007 movie. The 63-year-old made an appearance in the animation, which also foresaw Donald Trump becoming president and The Shard being built. [...] The Cast Away star said they suffered from aches, chills, and colds. And while the theory might seem a bit far-fetched, fans think it is The Simpsons’ way of predicting Tom’s isolation. It wasn’t long ago that The Simpsons fans believed that the show predicted the outbreak in 1993. Taking to social media one fan tweeted: ‘The Simpsons predicted The Coronavirus and Tom Hanks self-quarantine in 2 separate episodes? That show has predicted so many things!’ [...]</p> <p>https://metro.co.uk/2020/03/12/simpsons-fans-convinced-movie-predicted-tom-hanks-getting-coronavirus-2007-12390330/ (accessed on 4 January 2022)</p>
0.51	Agree	<p>‘The Simpsons’ predicted the coronavirus outbreak over 20 years ago. The animated prophecies of “The Simpsons” have long been documented by fans of the series. Now in its 31st year, the cartoon created by Matt Groening predicted many a world-altering event long before they took place, including Donald Trump’s presidency, Greece’s economic meltdown and the underdog American Olympic curling team besting the Swedes. [...] Speaking of “The Simpsons” predictions in general, Oakley said, “It’s mainly just coincidence because the episodes are so old that history repeats itself”. [...]</p> <p>https://nypost.com/2020/03/27/the-simpsons-predicted-the-coronavirus-outbreak-over-20-years-ago/ (accessed on 4 January 2022)</p>

Table 5. Cont.

		Example 3	
0.49	Discuss	[. . .] Earlier this year, a 1993 episode called Marge in Chains made the rounds on social media. That’s the episode with the “Osaka Flu” and the hornets. Bill Oakley didn’t want people to use their show for nefarious purposes on social media. “I don’t like it being used for nefarious purposes”, Oakley told The Hollywood Reporter. “The idea that anyone misappropriates it to make coronavirus seem like an Asian plot is terrible. In terms of trying to place blame on Asia—I think that is gross. I believe the most antecedent to (Osaka Flu) was the Hong Kong flu of 1968. It was just supposed to be a quick joke about how the flu got here”. [. . .]	https://comicbook.com/tv-shows/news/simpsons-classic-clip-fan-shares-wondering-will-this-horrible-year-end/ (accessed on 4 January 2022)

6. Discussion

According to the results presented in the previous section, it is evident that the stance classification is the component that negatively affects the overall performance the most. This happens because it is not able to properly generalize on unseen data. The good performances on the validation set (the accuracy was measured on a different split of the FNC-1 dataset from the one used for training), suggest that the problem is not the model capacity, but the FNC-1 dataset. Probably the FNC-1 dataset, used to train the stance classification model and the CoVID19-FNIR, used to validate the entire system on COVID-19 related fake news, have different language statistics and the stance classification model is not able to generalize on the second dataset. Hence, if the problem is not the capacity of the model, the only solution left is to expand the dataset, that is a very time-consuming task and it is out of the scope of this work. However, it is worth considering that the proposed method is particularly well-suited to a continuous learning implementation. For example the news corpus about coronavirus may evolve during the course of the epidemic and doc2vec can be fine-tuned periodically on the fresh news. The architecture of doc2vec allows us also to introduce new terms in the vocabulary. This can be achieved by adding a new random entry in the embedding matrix and let the model learn the proper weights. The new entry could also be obtained averaging similar words, in order to provide a smart initialization. For example, assume that a new vaccine for the COVID-19 is introduced, its embedding can be obtained averaging the embeddings of the other vaccines, in this way the new entry is ready to use and it will receive just small updates during fine-tuning. The same considerations can be made for news items that are too old and therefore do not present the latest information. The top-k similar documents indeed are obtained computing each time the cosine-similarity between the embedding of the query and the embeddings of the relevant documents, hence the list of relevant documents can change dynamically. Most important is the fact that the model could learn continuously from the user feedback. We already improved the results of the stance classification providing to the stance network manually labeled data. This suggests that this strategy can be expanded in the following way: when the system provides the list of similar documents categorized into the three categories *Agree*, *Disagree*, *Discuss*, the user is asked to give a feedback assigning the right label, including also the *Unrelated* label (to make this work it could be necessary to return also the unrelated documents, in order to give a feedback also on the false positives from this class). Then, the labels assigned by the user are used to fine-tune the stance network. We believe that the continuous learning implementation of the proposed model could considerably improve the performances.

7. Conclusions

In this work we presented a system that is able to find documents that are similar to a query. We also showed that the results of the stance classification, even though they are not so reliable to automatically classify fake news, can be used to further improve the

relevance of the content presented to the user and to better organize it. This allows the user to have at hand all the information needed to judge for themselves if the query can be considered true or fake. We have demonstrated through examples that the system supplies valid answers. In addition, we have outlined guidelines for future refinements, because we believe that the results that we obtained underestimate the real power of this method, that can be unlocked inserting the user feedback into the equation.

Author Contributions: All authors have equally contributed to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work has been partially supported by the Hermes-WIRED project within the Large Research Projects grant framework 2020 funded by Sapienza University of Rome.

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

References

- Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
- Gelfert, A. Fake news: A definition. *Informal Log.* **2018**, *38*, 84–117. [CrossRef]
- Gallè, F.; Veshi, A.; Sabella, E.A.; Çitozi, M.; Da Molin, G.; Ferracuti, S.; Liguori, G.; Orsi, G.B.; Napoli, C.; Napoli, C. Awareness and Behaviors Regarding COVID-19 among Albanian Undergraduates. *Behav. Sci.* **2021**, *11*, 45. [CrossRef] [PubMed]
- Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef] [PubMed]
- Oshikawa, R.; Qian, J.; Wang, W.Y. A survey on natural language processing for fake news detection. *arXiv* **2018**, arXiv:1811.00770.
- Wang, W.Y. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.00648.
- Long, Y. Fake news detection through multi-perspective speaker profiles. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Taipei, Taiwan, 27 November–1 December 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 252–256.
- Pham, T.T. A Study on Deep Learning for Fake News Detection. 2018. Available online: <https://dSPACE.jaist.ac.jp/dSPACE/bitstream/10119/15196/3/paper.pdf>(accessed on 4 January 2022).
- Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
- Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
- Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SigKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
- 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]
- Dungs, S.; Aker, A.; Fuhr, N.; Bontcheva, K. Can rumour stance alone predict veracity? In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 3360–3370.
- Tacchini, E.; Ballarin, G.; Della Vedova, M.L.; Moret, S.; de Alfaro, L. Some like it hoax: Automated fake news detection in social networks. *arXiv* **2017**, arXiv:1704.07506.
- Gupta, M.; Zhao, P.; Han, J. Evaluating event credibility on twitter. In Proceedings of the 2012 SIAM International Conference on Data Mining, California, CA, USA, 25 January 2012; pp. 153–164.
- Jin, Z.; Cao, J.; Jiang, Y.G.; Zhang, Y. News credibility evaluation on microblog with a hierarchical propagation model. In Proceedings of the 2014 IEEE International Conference on Data Mining, Shenzhen, China, 14–17 December 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 230–239.
- Jin, Z.; Cao, J.; Zhang, Y.; Luo, J. News verification by exploiting conflicting social viewpoints in microblogs. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
- Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. Defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 395–405.
- Ferreira, W.; Vlachos, A. Emergent: A novel data-set for stance classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1163–1168.

20. Yufeng. BBC Articles Fulltext and Category. Available online: <https://www.kaggle.com/yufengdev/bbc-fulltext-and-category/code> (accessed on 4 January 2022).
21. Byron Galbraith, D.R. Fake News Challenge FNC-1. Available online: <http://www.fakenewschallenge.org/> (accessed on 4 January 2022).
22. Dietterich, T.G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
23. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [[CrossRef](#)]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [[CrossRef](#)] [[PubMed](#)]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
27. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, PMLR, Beijing, China, 2–24 June 2014; pp. 1188–1196.
28. Lau, J.H.; Baldwin, T. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv* **2016**, arXiv:1607.05368.
29. Dai, A.M.; Olah, C.; Le, Q.V. Document embedding with paragraph vectors. *arXiv* **2015**, arXiv:1507.07998.
30. Rajendran, G.; Chitturi, B.; Poornachandran, P. Stance-in-depth deep neural approach to stance classification. *Procedia Comput. Sci.* **2018**, *132*, 1646–1653. [[CrossRef](#)]
31. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:cs.CL/1408.5882.
32. Wei, W.; Zhang, X.; Liu, X.; Chen, W.; Wang, T. pkudblab at SemEval-2016 Task 6: A Specific Convolutional Neural Network System for Effective Stance Detection. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 384–388. [[CrossRef](#)]
33. Julio, A.; Saenz, S.R.K.G.; Shukla, D. CoVID-19 Fake News Infodemic Research (CoVID19-FNIR) Dataset. 2020. Available online: <https://iee-dataport.org/open-access/covid-19-fake-news-infodemic-research-dataset-covid19-fnir-dataset> (accessed on 4 January 2022).