**Fake news detection using machine learning**

Naste Teodora Ioana, ICA

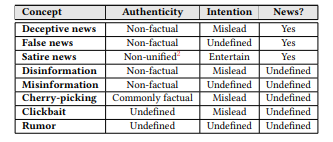
**Abstract**

Society has been heavily impacted by the effects of fake news, especially considering recent world events, like the COVID-19 pandemic. While a first step would be the education of the population in this matter and aiding the user with manual fact-checking already done by experts from various organisations, the volume and speed at which it spreads makes it so that it is impossible to keep up. This is why research into automatic fake news detection has been of a great interest these past years. With my research presented in this paper I want to highlight to importance of fake news detection and early fake news detection and methods that have been used to solve different concepts of fake news, the final goal in my mind being the implementation of a system inspired by the ones presented in here.

**Introduction**

Fake news have become a real issue in modern society when mostly everything has moved online where information is spread faster and cheaper than traditional news media such as newspapers and television. There have also been studies that have demonstrated that humans are only able to differentiate between legit and deceptive information with a mean accuracy around 55%-58% [11]. This was also proven by real-life events, like the COVID-19 pandemic, in which rumours of miraculous cures or potential non-existent threats have only managed to scare and confuse the people, inducing chaos in a time that was already unstable and uncertain. Of course, the fields in which fake news circulate are not reduced only to health conspiracies, but also in politics, economics and entertainment, so I believe it should come as a reflex to study this phenomena and try to combat it.

As one can tell, this is not an easy task and one of the issues is deciding what exactly are fake news. The term can be defined as information that it’s not true and it’s spread with a malicious intention, deception, rumours, clickbait or even satires and memes. Minyi Zhou and Reza Zafarani gather substantial information in their survey [11] on fake news and manage to discover 3 that unites these concepts: authenticity, intention and whether the information is news. Table 1 presents how concepts related to fake news are associated with each of these 3 characteristics. It is important to mention that even if a piece of information is not necessarily considered actually news that does not mean that it cannot manipulate the public’s opinion and that it doesn’t need fact-checking as much as any other piece of news. On the contrary, the difference is just in the method in which fact-checking is done, because different types of text have different features.



**Table 1. Comparison between concepts related to fake news**

This paper presents a survey of my research on the topic of automatic fake news detection and fact-checking, covering a large series of topic and methods. First section starts with a short theoretical explanation of the models and features that are mostly used for this task. Then in the second section, some of the most relevant datasets are presented, based on their topics. I felt this was necessarily because only recently experts have started to put together such labelled datasets. The third section contains state-of-the-art examples which are presented and explained briefly for different concepts of fake news. Lastly, a short description of the direction in which my future work will go.

1. **Theoretical aspects**

I believe that, after understanding a task definition, it is always important to understand the tools we have at hand before getting into work. That’s why in this chapter I will only briefly present the theoretical aspects of the models and features that have been used in the past for fake news detection.

* 1. **Linguistic features**

When working with machine learning and text , linguistic features are often used to better represent what we want the model to capture. For fake news detection , where there are many overtones to the text, these features should be given more attention as they can outline the style in which text is written. As it is said in [11], there are 4 levels at which these features are: lexicon, syntax, discourse and semantic.

At lexicon level, a good model to use is Bag Of Words, which represents a text or document just after its words’ frequency, ignoring their order. In the next sections it will be described how this can help to differentiate between fake and true news. To get a better view at the syntax of a sentence using a Part-Of-Speech tagger can help to get some insight, because studies show that fake news contain more adjective and nouns, while real news have more wh-words, prepositions and verbs [14].

Another method that helps fact checking is comparing a new text with one that’s already labelled, by calculating similarity between to texts. Using the TF-IDF index for similarity has been used several time with great results. This measure combines the term frequency, which is defined as the number of documents which contain a term *t* , with inverse document frequency, which helps rare words to rank higher, because the specific words for a topic are the ones that actually make the difference.

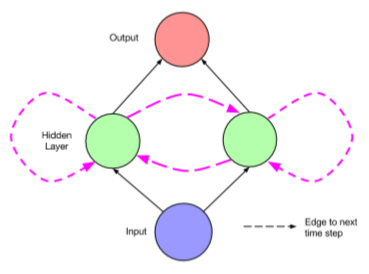
* 1. **Artificial Neural networks**

The model of the artificial neural networks (ANN) is inspired by the structure of the human brain. It’s formed by a big number of artificial neurons, also called units or nodes, united by edges. Each unit *j*  has an associated activation function and each edge has a weight , which represents the cost from unit *j* to the unit *j’*.

The classic neural network is capable of approximating a non-linear function. These networks learn using error propagation. The values calculated by the nodes is sent to the output nodes which compare them with the expected results. Each output node then returns an error. The purpose is to minimize this error. The simplest method is to distribute the error from one output node to all the other nodes connected with it , proportionally with the nodes’ weights, till the input nodes and each edge will update its weight and a new error will be calculated.

1.2.1 Recurrent neural networks

In contrast to traditional neural networks, the recurrent ones can have cycles, introducing the notion of time to the model. Thus, at a time *t* , the nodes with a recurrent edge get information from the current input and from the previous states . A simple recurrent neural network (RNN) is presented in Figure 1.2.1. [1] Moreover, RNNs are able to handle variable-length input, making them suitable candidates for time series predictions.



***Figure 1.2.1. Recurrent neural network***

The training for a RNN can be difficult. The main problems that can appear are vanishing and exploding gradients which can occur during error propagation. Which of these two events may occur depends on the weight of the edges and on the activation function. For example, if the activation function is a sigmoid one, the values risk to get too close to 0 and the network won’t learn. In this case, we deal with vanishing gradient. If he activation function is a ReLU max(1,x), then it’s easy to imagine the values growing too high and the exploding gradient might appear.

1.2.2 Long Short-Term memory

To solve the problems of the vanishing and exploding gradients that appear during training of a standard RNN, the Long Short-Term Memory neural network was proposed in 1997 [12] with an innovative gating system that introduces an intermediate memory space via the memory cell.

The gating system consists of 3 gates: the input, forget and output gates. The input gate, , decides how much of the information processed by the input node, , should be modified. The forget gate, helps to sort out the irrelevant information and the internal state, , is calculated based on the results from the previous 2 gates. Then, the final activation is composed with the help of the output gate, , which decides how much of the information should be passed to the next units. This process can be viewed in the next equations:

1.2.3 Gated Recurrent Unit

The gated recurrent unit neural network is a type of recurrent neural networks proposed recently which use a gating system to solve the vanishing gradient problem similar to the LSTM. It was first proposed in 2014 by Kyunghyun Cho et al [13] as a simplified version of the aforementioned RNN.

The gating mechanism is composed by 2 gates: the reset and the update gates. The reset get eliminated the past information that is considered to be irrelevant, thus helping the current unit memory at moment *t*  to have only the important information stored. The update gate helps to decide how much of the past information to be passed along to the future. The final activation of the unit is composed using the local memory and the results of the update gate and are passed to the next unit. Mathematically, this process consists of the next 4 equations:

1. **Datasets**

Besides knowing the models we have at hand for possible solution of the task of fake news detection, data is also one of the most important factors in making sure that the model is performing well and correctly. In the past years there has been interest in offering access to researchers to public sources of labelled data. In this chapter I will present a few of the databases that have been created for this task, classified based on the topics they cover.

2.1 Political statements

A well funded benchmark dataset was proposed by Wang in 2017 [10], called LIAR. Due to the lack of a well put together dataset for fake news detection, Wang managed to extract over 12.000 short statements from PolitiFact[[1]](#footnote-1), an American site for expert-based manual fact checking. Each entry in the data set is formed from the statement, speaker, context/topic, label and an explanation for the label assigned. Each statement is labelled with one of the following 6 rating labels: *pants-fire, false, barely-true, half-true, mostly-true* and *true.* The data is well balanced, including speakers from different political parties with a rich amount of meta-data for the speaker, such as: current job, home state or credit history and , also, statements from different sources (news, interviews, social media etc.).

2.2 Social media posts

The amount of data and the speed at which data is created are significantly higher on social media than any other platform, so there is plenty data to fetch. One dataset that has the investigation of rumours in social media at core is CREDBANK [4]. The dataset is composed of over 60 million tweets referring 1049 different real-word events. What makes this dataset interesting is the way it was labelled. Instead of going for an expert opinion, all the tweets in their pipeline were annotated by regular individuals to get a better insight at how users’ judgement can be influenced by social media. The labels are based on a scale going from *Certainly Inaccurate* (-2) to *Certainly Accurate* (+2) .

Another important repository meant for studying the patterns of fake news in social media is FakeNewsNet created by Kai Shu et al. [7]. Besides other previous datasets, FakeNewsNet incorporates extra meta-data (context and spatiotemporal data) that can help researching the diffusion of fake news in early stages which is a crucial aspect in social media where misinformation tends to spread faster than true information and it’s important to cut the thread earlier as possible. The ground truth for the datasets have been collected from well known fact checking sites: PolitiFact and GossipCop[[2]](#footnote-2) (focuses on entertainment news). The labels are the ones used by the experts on the respective sites. After that, posts related to the news extracted have been gather from twitter. Besides the posts itself, it contains information about how the users responded to the posts in terms of likes, comments and reposts.

2.3 COVID-19 claims

The pandemic that hit the world in 2020 was a global event that easily showed the impact that fake news can have and how fast it spreads. The safety of people could be endangered by some of the rumours circulating about possible treatments and symptoms of COVID -19. In this scope, experts have gathered in trying to combat this misinfodemic and as a result, several datasets that focus on topic related to COVID -19 and the pandemic have been created. One such dataset was released by Patwa Parth et al [5]. The dataset consists of over 10.700 social media posts and articles with topics regarding COVID-19 which have been manually labelled as real or fake. The fakes news have been collected from various social media posts that have been verified as fake and also from some fact checking websites (PolitiFact , Snopes[[3]](#footnote-3) etc.). The real news have also been gathered from verified social media posts made by official organisations or government accounts (World Health Organisation, Centers of Disease Control and Prevention etc.).

Another dataset that I think it’s worth mentioning is the one offered by Poynter[[4]](#footnote-4). The experts from this organisation (which is also responsible with PolitiFact) have fact checked and labelled thousands of claims regarding COVID-19 from all over the world. Their annotation system consist of many labels such as: *True, False, Mostly-true, Mostly-False, Misleading, Manipulated* etc. However, it is very easy to filter between these labels and the data is very easy to collect using web scrapping. Besides the claim and the label, each piece of data comes with an explanation and also a link to a verified article that confirms the tag. One more thing that should be taken into account is that it mainly focuses on claims that are inaccurate and there is a gap between the number of claims tagged as *True* and *False* (with their variations).

2.4 Mixed

There are many topics in which fact checks is needed, even if the main focus is on politics, entertainment and health care, which is very reasonable. However, generic fact checking can be a more difficult task considering the amount of different knowledge needed and the rate of getting new information that has never been seen before rises. FEVER [9] is a dataset introduced by James Throne et al in 2018. It consists of over 185.000 claims generated from Wikipedia on different topics. The data was collected by altering sentences from the first paragraph of Wikipedia articles, then these were manually labelled with one of the tags SUPPORTED, REFUTED or NOTENOUGHINFO.

Veronica Perez-Rosas et al.[[5]](#footnote-5) have released a dataset for automatic fake news detection on Kaggle. It contains articles from various mainstream news websites from the US. The topic of the news are grouped in one of the categories: technology, education, business, sports, politics and entertainment. Each article is marked with either *false* or *real.* Of course, this is not the only dataset on Kaggle that addresses fake news detection. Another one would be the official dataset from the Fake News Detection Challenge[[6]](#footnote-6) which contains articles from various topics labelled with 0 (False) or 1 (True).

A special mention on this topic should be also made towards web fact checking sites which offer nicely organized labelled claims and articles. Some of the most noticeable ones are: Poynter, Snopes, The Washington Post’s fact checker[[7]](#footnote-7), The Journal fact check[[8]](#footnote-8) etc.

1. **Related work**

The raise in popularity of social media has massively increased the amount of user-generated information that can spread uncontrollably throughout the web. This heavy amount of data generated in real-time is impossible to be filtered and checked manually for veracity. So, there has been a more interest to research automatic ways of detecting false information. However, considering the complexity of what can be considered fake news and the various forms in which this can be spread in, the task of simply determining the authenticity of the information can be very hard and complex.

* 1. **Stance detection**

Realising the complexity that fake news detection consists of, as a first step to tackle the task, in 2017 the Fake News Challenge[[9]](#footnote-9) released a contest for stance detection. Given two pieces of text, the purpose of the task was to estimate the relative perspective or stance of a text to the other on a given topic, claim or issue. The dataset for the challenge consisted in pairs of news article headline and its body and the contestants had to estimate the stance of the body towards the headline. The stance was represented by one of the following labels: *agree, disagree, discuss* orunrelated. There were many entries for this challenge, each with different types of models, however the ones that got into top 3 used rather simple methods.

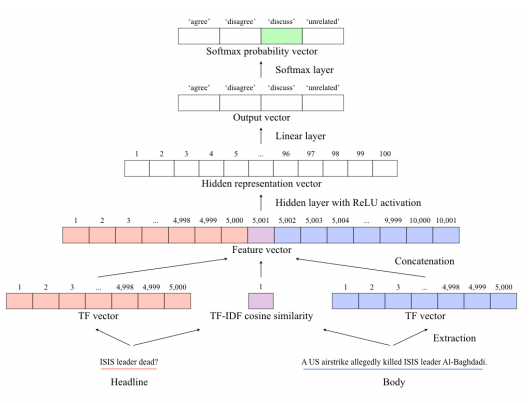
One of the approaches that got into top 3 was proposed by Benjamin Riedel et al. [2]. They probably had the most simple method, yet managed to get a considerable high accuracy.

**Methodology**

The method proposed consists of passing lexical and similarity features through a multi-layer perceptron with one hidden layer. The text was represented by a simple bag-of-words and the features considered were the following:

* Term frequency (TF) vector of the headline
* Term frequency (TF) vector of the body
* Cosine similarity between the TF-IDF vectors of the body and headline

The MLP classifier had 100 units on its single hidden layer. The rectified linear unit (ReLU) was used as the activation function and a softmax function on the output layer. In figure 3.1.1 a schema of the system is presented.

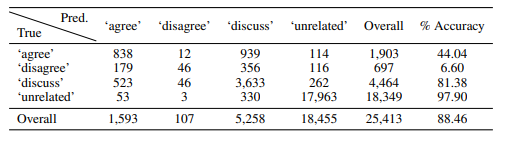


**Figure 3.1.1 Schematic diagram of the stance detection system**

A dropout and the Adam optimiser were also applied to avoid the gradient descent issue. The parameters for the model were learnt using a random search and cross-validation.

**Results**

The proposed method reached an accuracy of 81.72% which took the team on the third position at the end of the challenge. Results on different labels are detailed in Table 3.1.1., where it can be observed that the system performed best on text labelled with “agree”.



**Table 3.1.1 Confusion matrix for the stance detection system**

* 1. **Content based detection**

When talking about fake news detection, the first concept that pops into anyone’s mind is to check the information (content) from a piece of news and assign if it is true or false, without necessarily taking into account the intention of the creator.

Such approach has been taken by Souvick Ghosh an Chirag Shah [8] in their research paper *Toward automatic fake news classification*. Inspired by one of the top 3 teams from Fake News Challenge 2017 they wanted to see if combining information retrieval with stance detection and deep learning, they will be able to obtain good results on labelling text as ‘Fake’, ‘Legit’ or ‘Suspicous’.

**Dataset**

Because of a lack of benchmark datasets , the researchers used multiple datasets which they combined into 2 categories:

* Type 1, formed from LIAR, Kaggle’s fake News dataset and Fake News Challenge dataset, is a set of short texts with the length between 70 and 150 characters, including statements and article headlines.
* Type 2, formed from University of Washington Fake News Dataset, consists of much larger texts with the length between 400 to 700 words.

**Methodology**

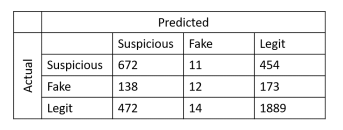
A submodular approach has been used. The first submodule was assigned to detect the veracity of the text using information retrieval, after previously constructing a knowledge base. First, documents referring the claim are retrieved from the knowledge base. For this the TF-IDF method was used, combined with a other advanced algorithms (BM25, Vector Space model, Language model) to improve the performance. After that, for each document retrieved the stance of the claim towards the document is determined. The second step has the task of classifying the claims into one of the 3 categories mentioned above. For that , they used a simple Feed Forward Neural Network that gets as input the bags-of-words vectors for the claim and the document, as well as the cosine similarity between these vectors. For training, the Type 1 dataset was used, as it is more focused on fact-based differences.

The second submodule is concerned with finding patterns in the writing style of the claim. It has been observed that fake news tend to be more aggressive and depict stronger emotions in order to manipulate the audience [11]. For this task, a bidirectional LSTM has been used, as these networks are known for their capabilities to process long sentences. This submodule was trained using the Type 2 dataset, as it has more richness in style compared to the short texts from Type 1.

Finally, results from both submodules are combined based on a voting system with the help of a weighted average.

**Results**

The prediction submodule managed to obtain an accuracy 67.1% for the ternary classification and of 72,12% for the binary classification, while the style based submodule obtained an accuracy of 81.83%. After combining the two submodules, they obtained a total accuracy of 82.4%. The experimental results for the prediction on FNC Dataset can be seen in Table 3.2.1.

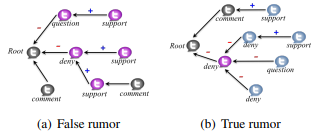


**Table 3.2.1 Confusion matrix for FNC Dataset**

* 1. **Fake news in social media**

Social media is the main environment in which fake news are being created and spread at an alarming rate. However, this makes space to research different aspects of fake news, like the propagation pattern and the users’ response via comments or reposts, with the intention to have an automatic method of spotting misinformation in its early stages of spreading.

In [15], have proposed a tree-structured recursive neural network approach that analyses how a rumour is spread and what’s the users’ stance relative to it. They started this researched based on observations they have made that suggest that if a user denies a post with a fake rumour, it tends to trigger positive responses from other users, confirming the denial, while denying a true rumour triggers negative and confusing responses. This pattern can be seen in Figure 3.3.1.



**Figure 3.3.1 Propagation tree containing users’ stance towards the post**

**Methodology**

A rumour (claim) was defined as an original tweet together with all its relevant responsive tweets ordered chronologically. For these claims the task is to be classified into one of the following categories:  *non-rumour, false rumour, true rumour* and *unverified rumour.* To capture the information from different angles, two variants have been adopted:

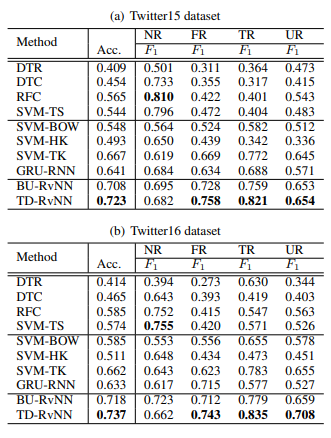
* Bottom-up RNN in which the direction of the edge points to the source, like in Figure 3.3.1
* Top-Down RNN which follows the natural propagation course in which the direction of an edge goes from the source to the user node that responded to that post

The purpose for each tree is to generate a feature vector that can be used as input to an extended version of the GRU neural network that Jing Ma et al. implemented. The difference between the two variants however is that, the bottom-up is going to start from the leaves and create a feature vector for each subtree and merging them together at the root creating the vector for the whole tree. On the top-down version, the creation is starting from the original post , resulting in feature vectors for each propagation line, since the final representation will be in the leaves of the tree. This enables a deeper insight into the complex propagation patterns that rumours can have on social media.

The error function used was the squared error. The parameters were updated using back-propagation through structure and also a gradient-based optimization was applied.

**Results**

The evaluation was done on two real-life Twitter datasets (Twitter15 and Twitter16) which each contain over 1000 propagations trees. To better present the advances of their work, Jing Ma et al. made an elaborate comparison with multiple state-of-the-art methods. The results on both sets can be seen in Table 3.3.1



**Table 3.3.1 Results comparison of rumour detection for each category**

From the table, it can be easily observed how the proposed model surpasses previous ones on most categories. Moreover, after obtaining such amazing results, further experiments were done to see the capabilities to detect rumours in early stages. They concluded that their method needs approximately 8 hours or about 90 tweets to correctly spot a fake rumours which is superior to the best baseline method which needs 36 hours and about 300 posts.

* 1. **Satirical news detection**

Satirical news bring another challenge to fake news detection, because their purpose is for entertainment and thus, it is harder to recognize the intention of the creator. Also, the satire can be subtle and so, users could easily misunderstood it as true news. The style of writing these kind of news it’s different than with other types of articles and it needs special investigation. Therefore, satirical news detection is researched separately from other news detection.

Considering this, Fan Yang et al. [16] have investigated satirical news and proposed a paragraph-level detection model that can deal with such subtle satire, but also highlighting the difference in features between document-level and paragraph-level. As they also pointed, satirical news are still considered fake news if the user believes them and spreads them as true information.

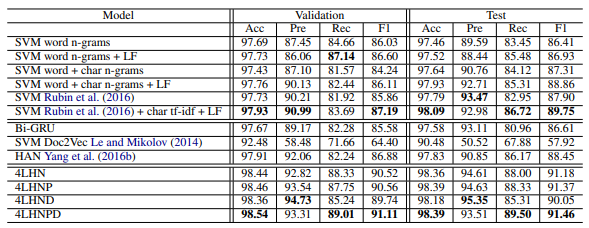
**Methodology**

To achieve their goal, they proposed a 4-level hierarchical neural network, following the hierarchy of character-word-paragraph-document.

Firstly, on the lowest level of the hierarchy, a convolutional neural network (CNN) is used to encode words representations from characters. Then these representations are sent one level up, where based the word embedding takes place. For encoding a Gated recurrent Unit (GRU) was implemented. The third level focuses on encoding at paragraph level, using a bidirectional GRU. Here, the attention mechanism is included because it was observed that not all paragraphs contain satirical cues and some paragraphs can be more relevant than others. For that, a satirical degree is being computed based on the paragraph representations and the results from the hidden states. The final document is the weighted sum of the satirical degree and the paragraph representations. On this level extra linguistics features are added. On the final level the final vector representation is computed via MLP and the prediction is done using a sigmoid activation.

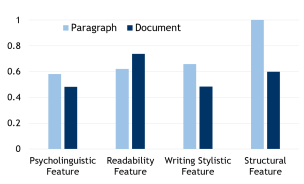
**Results**

For evaluation, over 16.000 satirical news were collected from various sites that focus on such content and over 160.000 true news from major news outlets. Since the data is not annotated, only the binary classification was taken into account. Moreover, different models were implemented to prove the efficiency of the method. These models as well as all the results can be seen in Table 3.4.1.



**Table 3.4.1 Results comparison for satirical news detection**

It can be observed that the model including both paragraph and document level features performs the best in almost all categories. The importance of the linguistic features used was also analysed which gives a very interesting insight into what features are more relevant when trying to detect satire. The results can be seen in Figure 3.4.1.



**Figure 3.4.1 Comparing the importance of the four features sets**

1. **Future work**

With this paper , I only wanted to highlight how vast and complex the field of fake news detection is by condensing some knowledge from my research. I think it is interesting to continue to research this field with a bit more focus on content based detection or veracity detection as this seems to have been the main focus especially in the last year of the pandemic. Also, I believe while binary classifications for this task can showcase very good results, that it might be too including and missing some textual hues, for example when parts of the truth are omitted for the purpose of misleading individuals. So, a comparison between binary and multi-tag classification would be an interesting research topic. As for my experiments, they are also focusing on content based classification with a combination of information retrieval and machine learning, inspired by some of the papers presented which I think has a lot of potential for future real-life applications for fact-checking articles even before they are posted with little to none of human assistance needed.

**Conclusion**

In this paper, I made an overall survey for the fake news detection task, highlighting it’s complexity and importance. Fake news has a big impact on society, especially in the era of social media where information can spread easy and fast. However, it is still hard to have a concrete definition of what exactly it’s considered fake news, hence the sub-fields of the task. Besides verifying the authenticity of the content, it is also relevant to analyse the style, language, intention of the creator, as well as propagation patterns. It is important to remember that fake news can come in many forms from social media posts, to news articles that can look legit, to even jokes that can seem harmless. So, it is of interest to research these topics and to pay them attention. Also, there are now more sources of labelled data available on many different topics which can help research in this direction.

**Bibliography**

[1] Lipton, Zachary C., John Berkowitz, and Charles Elkan. *A critical review of recurrent neural networks for sequence learning*. arXiv preprint arXiv:1506.00019 , 2015.

[2] Benjamin Riedel et al. , *A simple but tough-to-beat baseline for the Fake News Challenge stance detection task, 2017*

[3] Iftikhar Ahmad et al., *Fake News Detection Using Machine Learning Ensemble Methods*, 2020

[4] Mitra T. and Gilbert E., *Credbank: A large-scale social media corpus with associated credibility annotations*, 2015

[5] Patwa, Parth, et al., *Fighting an infodemic: Covid-19 fake news dataset*. , 2020.

[6] Rutvik Vijali et al., *Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking,* 2020

[7] Shu Kai et al., *FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media*, 2019

[8] Souvick Ghosh and Chirag Shah*, Toward Automatic Fake News Classification, 2017*

[9]Thorne James et al., *FEVER: a large-scale dataset for Fact Extraction and VERification*, 2018

[10] William Yang Wang, *“Liar, Liar Pants on fire”: A New Benchmark Dataset for Fake News Detection*, 2017

[11] Xinyi Zhou and Reza Zafarani, *A Survey of Fake News: Fundamental theories, detection methods and opportunities,* 2020

[12] Hochreiter, S., & Schmidhuber, J.. *Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.* doi:10.1162/neco.1997.9.8.1735 , 1997

[13] Cho, Kyunghyun et al. , *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation ,* 2014

[14] Kapusta Jozef et al. , *Comparisonof fake ad real news based on morphological analysis*, 2020

[15] Jing Ma, Wei Gao and Kam-Fai Wong, *Rumour Detection on Twitter with Tree-structured Recursive neural Networks*, 2018

[16] Fan Yang et al., *Satirical news detection and analysis using Attention Mechanism and Linguistic Features,* 2017

1. <https://www.politifact.com/> [↑](#footnote-ref-1)
2. <https://www.gossipcop.com/> [↑](#footnote-ref-2)
3. <https://www.snopes.com/> [↑](#footnote-ref-3)
4. <https://www.poynter.org/ifcn-covid-19-misinformation> [↑](#footnote-ref-4)
5. <https://www.kaggle.com/sumanthvrao/fakenewsdataset> [↑](#footnote-ref-5)
6. <https://www.kaggle.com/c/fakenewskdd2020/overview> [↑](#footnote-ref-6)
7. <https://www.washingtonpost.com/news/fact-checker> [↑](#footnote-ref-7)
8. <https://www.thejournal.ie/factcheck/news/> [↑](#footnote-ref-8)
9. <http://www.fakenewschallenge.org/> [↑](#footnote-ref-9)