Society and Economy 44 (2022) 2, 183–205
DOI: 10.1556/204.2021.00020

AKADÉMIAI KIADÓ

Fake news identification

PETER RACSKO[*]

Institute of Informatics, Corvinus University of Budapest, Budapest, Hungary

Received: November 8, 2020 • Revised manuscript received: September 10, 2021 • Accepted: October 11, 2021
Published online: November 11, 2021

ABSTRACT

Fake news, deceptive information, and conspiracy theories are part of our everyday life. It is really hard to distinguish between false and valid information. As contemporary people receive the majority of information from electronic publications, in many cases fake information can seriously harm people's health or economic status. This article will analyze the question of how up-to-date information technology can help detect false information. Our proposition is that today we do not have a perfect solution to identify fake news. There are quite a few methods employed for the discrimination of fake and valid information, but none of them is perfect. In our opinion, the reason is not in the weaknesses of the algorithms, but in the underlying human and social aspects.

KEYWORDS

fake news, Byzantine Generals, blockchain, machine learning, natural language processing, fact checking

1. INTRODUCTION

Anybody can create information in blogs, emails, social networks or online media. The information, however is hetereogenous due to the variety of interest and opinion of the sources. Politicians, influencers, and sometimes scientists publish inconsistent and contradictionary information on several topics. Cybercriminals use the Internet as effective and inexpensive way of distributing information serving their interests. There is no quality assurance on the Internet: anyone can post anything. In most cases, information found on the web has not been checked for accuracy.

[*] Corresponding author. E-mail: peter.racsko@uni-corvinus.hu

True and fake information appear on billions of nodes of the information network, online news channels, social networks, blogs, tweets, printed or broadcasted information sources. In the pre-internet era, it was easier to decide if the information was valid, thoughtfully checked and cross-checked encyclopedias, textbooks and prestigious newspapers served as the primary and authentic sources of information, people could rely on the information available in these sources (Vedder – Wachbroit 2003). The situation however has dramatically changed (Whitty – Joinson 2008). Today, as the internet became the primary and for many people the only source of information, the authority and righteousness of many sources has become increasingly questionable. Fake news is not only used for financial or political gain. Areas such as vaccination are also plagued by fake news (Susarla 2021). But which news *is* fake? How to identify fake news? Some people use the term fake news simply for facts or statements they do not agree with.

Let us try to give a working definition of fake news. It is intuitively clear what the term fake news means. It is necessary however to provide a more or less exact definition to this soft object or at least explain what we mean by fake news in a given context.

In dictionaries, news are explained as materials published traditionally in newspapers, today mostly on the internet. Fake means counterfeit.

Watson (2018) refers the following definition: fake news are "deliberately and strategically constructed lies that are presented as news articles and are intended to mislead the public". This has been a very exact definition during the history from the Ancient Roman Empire to the 20th century (MacDonald 2017). Today, however, we live in the Information Age which has a new name—the post-truth era (Wendling 2018; Harari 2018; Lewandowsky et al. 2017).

The Oxford Dictionary defines post-truth as "an adjective [. . .] in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief." Today many people propagate false information because they believe it is true. For practical purposes, in this paper we use the term "true news" for evidence based facts and false information or "fake news" for the facts which are not proven by evidence or they are proven false by evidence.

This question eventually proved to be a question of life or death in the Coronavirus pandemic, when some people took harmful drugs following misinformation (Coleman 2020.)

Social sites spread a lot of invalid information (Del Vicario et al. 2016). Facebook, Twitter, and online newspapers have been identified as the best platforms for spreading, but also monitoring misinformation and dispelling rumors, stigma, and conspiracy theories (Islam et al. 2020). Many people however have been considering the "official", government supported information as truthful, but in several instances even high level government officials spread rumors on conspiracies of other states or the unproven effectiveness of specific drugs. People without high level medical training cannot decide if a piece of information is valid or not.

How can individual persons, groups and the society today decide on the validity of a piece of information if they have no academic background in the specific area? Evidence based thinking (Shargel – Twiss 2019)? The problem is the same as before, why should we believe, that a piece of information broadcasted on the internet is based on real evidence or perhaps the evidence itself is fake? In our world with a wide variety of contradicting information, finding and understanding reliable evidence in most cases is beyond the capabilities of people, who have no specific knowledge. E.g. how could a person without studying quantum physics accept or reject a fact from quantum physics? Or a non-biologist accept the detailed facts of evolution?

The internet is a vast warehouse of false information e.g. quite a few conspiracy theories, such as chemtrails, health risks of 5G communication, false consequences of immunization and fake news on political or business issues. In many cases, social networks are the primary platforms for advertising false information. While social networks now try to control the quality of information posted on their sites, but they do not want to restrict the freedom of speech, even if the posted information is obviously false (Mosseri 2017).

During the long history of mankind, fake news was always present (Burkhradt 2017), but the ground truth was accepted by the vast majority of the people, because they either directly experienced the facts, as e.g. objects fall down, or because the fact was advertized by reliable people. Of course, people with high reputation in the past and even today may propagate false information due to lack of knowledge (for example the denial of the heliocentric cosmos in the medieval ages). Most people could not directly check if the Earth was flat or spheric, but they believed the fact as it was claimed by people whom they trusted because of their qualification, position, etc. Vast majority of people trusted the authority of scientists and religious or political leaders. It was a consensus of the majority, reached by accepted procedures, e.g. printed books, lessons given by teachers, lectures held by scholars. These procedures were the tools, the protocols of the consensus.

Today however, the role of well-known and prestigious personalities guiding people in the identification of the validity of a piece of information is decreasing, and partly dissolved in the ocean of "influencers" and opinion leaders, who spread their own views about the validity of a piece of information (Altay et al. 2020).

How can an ordinary person orientate herself in the jungle of contradicting opinions?

In this paper we demonstrate that this problem is not yet solved by computer algorithms and the human element is essential in the identification of false information.

Fake news might appear in the form of internet memes. How to identify them? Internet memes are digital objects e.g. photos, videos, gif files that are circulated on the internet in a "memetic manner". Memes usually carry secondary meaning, irony, sarcasm, or even more complex visual interpretation. The relative impact of memes is far greater than their volume and it would be nice to provide validity check to memes too, but in most cases their hidden message cannot be interpreted by algorithms, this task requires the human brain.

2. METHODS

In our paper – following a short theoretical background of reaching a consensus on the validity of information – we describe a few widely accepted models, algorithms and methods, and demonstrate that none of them is perfect and human perception is a fundamental factor of the identification of false information and fake news. We also demonstrate that the consensus and the practical process of an agreement in many cases works only in a closed community.

Further on, for the sake of simplicity we will take a situation when there are two contradictory statements about a given fact presented on the internet, true or false (e.g. the Earth is flat vs. the Earth is spherical).

We describe a method of finding a consensus between a set of participants – people or computer nodes, representing people or groups of people. A widely accepted and used theory of consensus on a particular piece of information is very well described by the Byzantine Generals' problem, and associated blockchains. We will demonstrate that blockchains can provide a consensus on the validity of information only if restrictions are applied to the participant or nodes. But the required conditions are usually not satisfied on the broad internet.

Then we present the state of the art of natural language processing (NLP) and machine learning (ML) in identifying false information. NLP and ML models however are highly dependent on the training data used for the model creation. The data analyst has a fundamental role in the characterisation of true or false information when she collects and feeds the training data.

As a human-based fact identification method, fact-checking services are available for the subscribers. An overview and characterisation of human based fact-checking services is given as the ultimate, but expensive tools for identifying fake news.

3. THE BYZANTINE GENERALS' PROBLEM

The Byzantine Generals' problem (Lamport et al. 1982) is a metaphor. The problem was originally called Chinese Generals' Problem, later Albanian Generals' Problem, but computer scientists at last accepted a less contemporary name. The fictitious problem of the Byzantine Generals is to synchronize the time of attack of the separated armies of the Generals, while the communication is asynchronous, and they do not trust each other. It is also known as source congruency and widely used in computer science for the description of a situation where the elements of a distributed system may fail and the information whether an element has failed or not may be fake. The problem is the development of a common strategy that helps to avoid system failure when the parties do not trust each other and they do not know if an element failed or not. The model is also known in cryptography as Byzantine Agreement, it is a situation when "honest" processors try to agree on something, but "adversarial" processors try to prevent the agreement and the participants do not know which processors are "honest" and which ones are "adversarial".

The fundamental question in all cases is how to design a method for reaching the agreement between the participants that ensures the truthfulness of an information with high probability even if the dishonest ones do not agree with the consensus.

It is reasonable to set up formal rules and processes for reaching an agreement. The rules and processes of negotiation between the Generals – methaphores for the nodes of a network – are integrated in a so-called consensus protocol. The consensus protocols are used in connection with blockchain models, the protocol is the method of agreement of the participants on the validity of a transaction (Racsko 2019). We will investigate the most popular and widely used consensus protocols and discuss if they can be used for the agreement of the validity of information. The problem of obtaining Byzantine consensus – a satisfactory consensus – was conceived and formalized by Lamport and co-authors (Lamport et al. 1982). The problem was called interactive consistency (Diamantopoulos et al. 2015). Interactive consistency is the equivalent of fault tolerance in computer systems, and existence of consensus protocols applied in blockchain technology.

Blockchain and its applications play a significant role in the economy as the fundamental technology for cryptocurrencies, contemporary logistic systems, smart contracts, etc. The mathematics behind blockchains were first described by Ralph Merkle and his construction was named the Merkle tree (Merkle 1982). It is a fundamental part of blockchain technology in applications, but another element, the formal description of the consensus procedure also must be associated with the Merkle tree for real applications. The blockchain itself is a theoretical proof of the validity of the data blocks but it is also a very important practical method of "writing" and "reading" the blockchain. This procedure is the consensus protocol for a given application, which is a specific feature of the application.

The development of hundreds of cryptocurrencies and many other applications (e.g. smart contracts) during the past years based on blockchain technology resulted in quite a few alternative consensus protocols (Bashir 2020).

In real world blockchain applications the consensus protocols consist of a set of rules deciding whether to accept a new transaction or block, if it is true or to reject if it is false. The set of rules is preliminary agreed by the participants and only by the participants, there are no general guidelines for the protocols.

We have millions of information pieces on the Internet. Is it possible to identify false ones using a consensus protocol based on the solution of the Byzantine Generals' Problem?

Below, we give a short overview of the mathematical results for solving the Byzantine Generals' Problem, the consensus protocols and their applications in blockchains, then we will outline a possible scenario for the identification of the fake news based on the blockchain technology and consensus protocols and review their characteristics.

The basic question is whether it is possible to develop and implement a consensus protocol for accepting or rejecting the validity of statements posted on the internet with high probability, or whether it is possible to develop a reliable and computationally feasible set of decision rules, which enables the user to decide if a statement is true or false. Can one replace the evidence based evaluation of the validity of a statement with a consensus-based agreement?

This situation can be formulated abstractly using the terms of the Byzantine Generals problem. Communicating only by messengers, the generals must agree if they will attack or not. However, a few generals – we can call them traitors – will try to obstruct the agreement with false messages. The problem is to find an algorithm to ensure that the loyal generals can reach an agreement (Lamport et al. 1982).

For further discussion we have to lay down some necessary features of the network:

i) The network has $\underline{n}$ nodes of which no more than m are faulty. It is not known, however, which nodes are faulty and consequently, the nodes do not know which other nodes are faulty. The nodes represent the generals, the faulty nodes are the traitors, and the valid ones are the loyal generals.

ii) The nodes have peer-to-peer communication.

iii) The communication network is complete (each pair of nodes can communicate with each other).

iv) Message integrity is guaranteed – the receiver of a message can check if a message was changed in the transfer or not.

v) A node cannot impersonate another node – the nodes do not necessarily reveal their identity, but cannot fake the identity of other nodes.

We call the final outcome of the exchange of messages with the features i) - v) an agreement, if:

- After a finite number of iterations all nodes select either a valid or invalid information.
- The valid information is identical for all valid nodes.

Lamport et al. (1982) provide the mathematical solution of the problem in two basic situations:

1. The communication is carried out with oral messages (OMs). OMs are the messages, whose content is under the full control of the sender, which is typical for computer networks.
2. The communication is based on Signed Messages (SMs). SMs restrict the communication in the following way:
   - All messages are signed by the sender.
   - The signature of a node with valid information cannot be forged, and breach of integrity of the messages can be detected.
   - All nodes can verify the authenticity of a node's signature.

A mathematical proof in the case of networks with the i) - v) features was given by Lamport et al. (1982), that with oral messages the condition $n > 3m+1$ is necessary and sufficient for the existence of the consensus protocol. The central part of this theorem is the proof of non-existence of a consensus protocol if there are 3 or fewer generals and one of them is a traitor. If there are at least 4 generals and the number of traitors is less than one third of all generals, the consensus protocol exists. The above consensus protocol by the way will not expose the traitors.

In the case of signed messages, traitors' signatures are not necessarily valid; they can even cooperate in sending false information. SMs in electronic communication are messages with electronic signature. Lamport et al. define the SM(m) consensus protocol, and prove that SM(m) gives a consensus if the number of traitors is not more than one third of the generals. SM(m) satisfies the interactive consistency conditions if there are only 3 generals and 1 of them is a traitor.

In both scenarios, if a general sends valid information to the other generals, the information is identical to that sent to all other generals, but if general sends invalid information to other generals, this is not necessarily identical for all addresses.

In distributed computing, the problem is how to reach consensus on the right information in a system where none of the network nodes can be trusted (Shi et al. 2021). The problem of distributed computing is how to build a fault tolerant network system if a part of the nodes is potentially faulty. We do not suppose synchronicity or reliance of known delay limitations in the communication between the nodes, because the internet is not synchronous and delays are not always predictable, but we allow arbitrary behaviour of the nodes (transmission of arbitrary information). Synchronicity cannot be supposed as in the case of malicious misinformation, the faulty information can be intentionally delayed. The Practical Byzantine Fault Tolerance (PBFT) algorithm presented in the paper of Castro and Liskov (1999) can reach consensus if the number of faulty nodes is less, than [(n-1)/ 3]. In case of only two or three nodes the algorithm does not provide a consensus. As the theoretical background was laid down, many practical protocols were developed with the aim of providing high performance. In 1999 Castro and Liskov published the PBFT algorithm that can manage Byzantine systems with a high performance.

In 2014 a solution for the fault tolerance methods was provided by Malkhi and Reiter (1998) also for transient faults. A fault is transient if a node is faulty in one step, but it can become non-faulty in the next step. We do not extend our research for transient faults, we suppose if a node is faulty, it remains faulty until the protocol stops. This assumption is not always applicable in fake news identification, as the nodes spreading fake news can change their content.

The interactive consistency protocols were developed as iterative algorithms. We will use the existing and well-studied blockchain technology for reaching the consensus in the Byzantine Generals' problem. Consensus in a practical application in blockchain technology is reached by a selected consensus protocol which is accepted by the participants of the application. Practical implementations of the technology, e.g. cryptocurrencies and smart contracts, always include a specific consensus protocol, which is an inseparable part of the application.

4. THE CONSENSUS PROTOCOLS

Let us start with the common features of the consensus protocols developed and used for the blockchain applications. The blockchain is a specific Merkle tree. Ralph Merkle patented the idea of the hash tree (or Merkle tree) in 1979 as a method for the verification of the validity of sequential or tree-like data structures (Merkle 1988).

Suppose we have a block of data, e.g. a set of payment transactions. The Merkle tree is a tree-graph, in which a label is attached to each node, leaf nodes are labeled with the hash of data blocks, non-leaf nodes are labeled with the hash of the labels of the child-leaves. Any standard hash algorithm can be applied. A Merkle tree allows a block of information (usually payment transactions) to be verified for accuracy efficiently and quickly. The concept of the Merkle tree is an important building block of blockchain technology applied in the cryptocurrency world.

The widely used form of a hash tree is a hash chain, where each non-leaf node has exactly one child node. Hash chains are used not only in cryptocurrency systems but in several areas of computer science for synchronization of data distributed over the network.

The hash chains are excellent tools for the proof of validity of the non-leaf nodes if the users of the chain previously agree on a set of rules that allow or deny the attachment of new leaves to the existing chain. The user community of the chain must have a consensus on the compliance with the rules. The consensus is reached via a previously set formal algorithm, called a consensus protocol. There exists a great variety of consensus protocols, let us shortly explain the most popular ones.

4.1. Proof of Work (PoW)

Dwork and Naor (1992) suggested the application of PoW to filter out junk mails. According to their idea, the sender uses a price function set by an independent pricing authority and a randomly chosen “shortcut” value (a.k.a. “nonce”). If one knows the shortcut, it is easy to compute the price function, if not, it is moderately hard. The pricing function is a hash function, where the email is hashed along with the receiver address and timestamp. The receiver checks the hash and accepts the mail only if the hash is correct.

The idea is that it is expensive for the spammer to send millions of emails, while verification is easy as the receiver knows the shortcut value. PoWs today have many applications, e.g. they are used to prevent double spending in the Bitcoin digital currency particularly. Bitcoin however, because of the high energy requirement of its consensus protocol has even been labelled an “environmental disaster” (Reiff 2021) as the implementation requires solving very hard computational problems and not a moderate one as suggested originally by Dworak and Naor.

Proof of works ensures that the minority of the participants cannot change the information. The weakness is the computational expense and the time required for the work. The PoW algorithm is not 100% tolerant to the Byzantine faults, but has proven to be one of the most secure and reliable implementations.

4.2. Proof of Stake (PoS)

The creator of the next block in the chain is chosen via a selection algorithm where the users with higher stakes (e.g. ownership of a large sum of cryptocurrency, or having a position in an organization, or other credentials) has a higher chance of getting the right to create the next block. The algorithm uses a random election process to select a node to be the creator of the next block. The chances also depend on a few factors as the staking age, and the node's wealth (Wenting et al. 2017).

In order to effectively control the network, a node would have to own a majority stake in the network. The strength of this protocol is cost-effectiveness, while the weakness is that the creators do not lose their stakes if they cheat, because they own the majority of stakes, unless the community penalizes them by some other external mechanism.

4.3. Proof of Space (PoSp)

PoSp is an alternative concept for PoWs, where a service requestor must dedicate a significant amount of disk space, rather than expensive processing time as in the PoW construction. The PoSp scheme is used in the currency system Spacecoin (Dziembowsky et al. 2013).

4.4. Proof of Burn (PoB)

Proof of burn (Karantias et al. 2017) is another alternative of consensus algorithm that addresses the high energy consumption issue of a PoW system. PoB is often called a PoW system without energy waste. It operates on the principle of allowing miners to "burn" virtual currency tokens. They are then granted the right to write blocks in proportion to the coins burnt. Iain Stewart, the inventor of the PoB algorithm, uses an analogy to describe the algorithm: burnt coins are like mining rigs. In this analogy, a miner burns their coins to buy a virtual mining rig that gives them the power to mine blocks. The more coins burned by the miner, the bigger their virtual mining rig will be. To burn the coins, miners send them to a verifiably un-spendable address. This process does not consume many resources (other than the burned coins) and ensures that the network remains active and agile. Depending upon the implementation, miners are allowed to burn the native currency or the currency of an alternate chain, such as Bitcoin. In exchange, they receive a reward in the native currency token of the blockchain.

4.5. Proof of Authority (PoA)

Proof of Authority (PoA) is a reputation-based consensus algorithm. The term was proposed in 2017 by Gavin Wood (Cappacioli 2019). In the PoA algorithm block, validators are not staking coins but their own reputation instead. PoA blockchains are secured by the nodes that are selected as trustworthy entities. The number of validators is limited, the participants acting as moderators are pre-approved. One of the main application area is logistics and supply chains, as companies can maintain privacy when participating in the blockchain system. Microsoft Azure uses the PoA for managing a multimember consortium. PoA is also used by the VeChain Thor blockchain, operated by the VeChain Foundation (2019) platform.

PoA is characterized by low computation power required to achieve network security and consensus integrity, and is controlled via built-in mechanisms preventing cheating. There are three types of stakeholders with different authority, the Authority Masternodes, Economic X

Nodes and Economic Nodes. The stakeholders can be individuals and organizations, but the Authority Masternodes have to be identified by the community. The new blocks are created by one of the Authority Masternodes. The actual creator of a block is selected by a random algorithm. The creators are rewarded for this activity. All Authority Masternodes have equal chances to create a block.

4.6. Proof of Elapsed Time (PoET)

Proof of Elapsed Time (PoET) is a consensus mechanism algorithm that is often used on the permission blockchain networks to decide the mining rights or the block winners on the network. Permission blockchain networks are those which require any prospective participant to identify themselves before they are allowed to join. Based on the principle of a fair lottery system where every single node is equally likely to be a winner, the PoET mechanism is based on spreading the chances of winning fairly across the largest possible number of network participants. The PoET concept was developed in 2016 by Intel Corporation (Chen et al. 2017).

4.7. Proof of Capacity (PoC)

Proof of Capacity emerged as one of the many alternative solutions to the problem of high energy consumption in PoW, the problem that promotes cryptocoin accumulation instead of spending as in PoS (Hayes 2021).

In the standard and commonly followed PoW consensus algorithm, the miners are constantly changing a number in the block header as fast as they can, aiming to find a correct hash value. The first miner to identify the correct hash value, and associated *nonce*, broadcasts that information to the network. Other miners validate and authenticate the transactions before moving on to work on the next block. Essentially, this approach works like a lottery system, where the miners keep changing the hash value to find the correct one. Proof of capacity however allows the mining devices (nodes) of the blockchain network the ability to use empty space on their hard drive to mine the available cryptocoins. Instead of repeatedly altering the numbers in the block header and repeated hashing for the solution value, PoC works by storing a list of possible solutions on the mining device's hard drive even before the mining activity starts.

The larger the hard drive, the more possible solution values one can store on the hard drive, the more chances a miner has to match the required hash value from his list, resulting in more chances to win the mining reward.

The above list of consensus protocols is far from complete. Some of them (e.g. PoW) are extensively studied for security and performance, while others have not been in the focus because of relatively low technical or economic interest.

4.8. The delegated Byzantine Fault Tolerance (dBFT) protocol

The term "Byzantine Fault" is derived from the Byzantine Generals' Problem, i.e. the faults are Byzantine. The dBFT (Zhang et al. 2021) algorithm is an iteration of the PoS protocol. It is used by cryptocurrencies EOS and Peercoin. The application of the protocol requires permanently connected nodes. Adding a new block to the blockchain is enabled by a consensus through proxy voting. A node can pick a "bookkeeper" it supports by voting. The selected group of bookkeepers, through the dBFT algorithm, reach a consensus and generate new blocks. Voting continues in real time. Achieving true consensus in a complicated task.

The primary application area of dBFT is distributed computer systems, where in a case of network crash the parties have to agree on the single component that failed. The problem is complex when many parties are involved in a network crash. Without singular consensus, it is difficult to determine which components failed.

The dBFT system ensures that if three quarters plus one of the network's nodes are correct, the faulty nodes are ignored when the final decision on the faulty node is made. The agreement is reached in final iteration steps. The dBFT provides a relatively high tolerance to bad actors.

5. IS FAKE NEWS IDENTIFICATION A BYZANTINE PROBLEM?

If three quarters plus one node agrees on a decision, and they do not change their opinion, then the problem is Byzantine. In a real situation when we have to decide if a piece of news is valid or not, it is hard to check if the conditions are met. We do not know the number of the nodes, publishing the news, consequently we cannot tell how much is three quarters are. It is also hard to fix the opinion of the valid nodes for a long time.

The consensus algorithms however, created for the blockchain systems, require the Byzantine feature, sometimes in an implicit way. If we want to apply one of the popular consensus algorithms to determine fake news, we should check if the problem is Byzantine or not. The techical requirements i) to v) are met in a normal internet environment, but the question is, if there exists a final exchange of messages that after a finite number of iterations all nodes select either a valid or invalid information and the valid information is identical for all "loyal" nodes which do not change their selection during the exchange. In an open internet environment it is hard to ensure this condition.

In cryptocurrency systems and other blockchain applications participants (even the malevolent ones) share a common interest, they usually do not challenge the existence of the currency, as they want to make a profit from having as much currency as possible, or they want to manipulate the exchange rate, or simply steal currency without undermining the basis of the system.

In case of publishing general information, only a part of the participants accepts the validity of the ground truth, while another part strongly believes in the validity of alternative scenarios, or wants to deceive malevolently other participants. Sometimes a part of the participants want to destroy the information exchange system. In a public information exchange environment, the open internet, there is no guarantee that the valid information is identical for all "loyal" nodes after a finite number of iterations and the number of loyal nodes n is greater than 3m+1, where n is the number of nodes who agree on the ground truth and m is the number of nodes broadcasting fake information. Several examples confirm this statement as we have experienced e.g. during the coronavirus pandemic.

The conclusion is that no public blockchain with the existing consensus protocols can support a generally acceptable verification system. As it was already mentioned, public cryptocurrencies can exist only because it is in the interest of the majority of participants to support the system. Even in the case of a 51% attack, it is essential for the attackers to save the system's integrity (Zimwara 2020) otherwise they cannot pocket the profit. A 51% attack is an attack on a blockchain by a group of miners who control more than 50% of the network's mining hash rate. Attackers with majority control of the network can interrupt the recording of new blocks by preventing other miners from completing blocks.

What can we say about permissioned or private blockchains with controlled access of participants? Private blockchains are based on access controls which restrict the set of people who can participate in the network. There are one or more entities which control the network and this leads to reliance on third-parties to transact. Let us analyze the different types of consensus protocols for private blockkchains.

Proof of Work is clearly not a good choice in a private network, as it requires a certain amount of resources from those who want to certify the information – perhaps without reward. Proof of Stake, Proof of Space, Proof of Burn, Proof of Elapsed Time, and Proof of Capacity have similar disadvantages.

The delegated Byzantine Fault Tolerance consensus protocol requires at least *(n-1)/3+1* of the network's nodes to be correct for reaching a final agreement. In case of verification of the validity of information, the community can reach an agreement only if at least *(n-1)/3+1* of the bookkeepers share the valid information and they do not change their opinion during the verification process. If all nodes can vote for all other nodes, we face the same problem, how to ensure that the number of righteous proxy nodes exceeds the minimum required for the Byzantine Agreement, otherwise the dBFT protocol might fail to reach the consensus.

Proof of Authority is reputation based. In fact, the traditional method of verification of the validity of information is also reputation based. People have been trusted in authorized organizations and other people with high reputation in their specific area. The Authority Masternodes are nominated by different types of community members, the enterprises, the developers, researchers, community contributors, and business development ambassadors. They must also have a minimum amount of VeChain tokens and their identity must be public to the community. The actual block is verified by a randomly selected Authority Masternode.

The question is, whether a community can use PoA for the verification of a given piece of information. The answer is definitely yes, but only if the community shares the same views, e.g. a belief in a flat Earth, or, to the contrary, the members rely on checked facts. And again, the initial setup of the system will determine if Proof of Authority consensus works or not. If someone joins the community with beliefs in scientific, fact-based information, she will stay with this community. But, if someone initially joins a community with beliefs in non-fact based, non-scientific information, as a result of continuous verifications she will not accept PoA and leave the community. This setup leads to subsets of network nodes with their own beliefs in truth with their own Authority Masternodes.

Why do not we apply simply the “wisdom of the crowd” approach? The idea of the power opinion of the crowd goes back to Aristotle, who is credited to agree with the idea that “many heads are better than one” (Landemore 2012).

Can we use the wisdom of the crowd to decide if a piece of information is true or not? We can apply the Byzantine theorem to the crowd. If the number of true individuals in the crowd exceeds 3m+1, then one can accept the opinion of the majority. But there are at least two major problems. The first one is that the Byzantine generals have to agree only on one particular decision, attack or not. On the internet, we have millions of news to be categorized to the true and the false class and maybe a set of people will satisfy the conditions of the Byzantine theorem in relation to one question, but not in others. The second issue is that the set of internet users is

Table 1. Comparison of characteristics of some consensus protocols

Name of the protocol	Finality of the protocol	Important features
PoW	Consensus is not final	Very expensive
PoS	Consensus is not final	Participants are not equal
PoSp	Consensus is not final	Participants are not equal
PoB	Consensus is not final	Participants are not equal
PoA	Consensus is final	Decision is made by selected participants
PoET	Consensus is final	Identification required
PoC	Consensus is final	Participants are not equal
dBTF	Vulnerable to faulty nodes>(n-1)/3	Voting by proxies

Source: author.

strongly divided. Most users look at the internet through a filter. The term “filter bubble” refers to the results of the algorithms that dictate what we encounter online (Pariser 2012). In most cases the wisdom of the crowd will depend on the filter bubble of the particular group of people (an actual example are the anti-vax versus pro-vax groups). The conclusion is that the wisdom of the crowd is not independent of the filter bubble of the particular group.

The conclusion on the use of blockchain technology for the identification of fake information on the internet is that the existing consensus protocols do not provide a solution to the identification of fake news in general (see Table 1). But if a person or organization joins a group with similar views or accepts the same organizational rules, it is possible to create a protocol for the verification of the information for the given group, e.g. the Proof of Authority protocol fits this goal, but it is not anonymous, as the Authority Masternodes are publicly identified.

In the blockchain setting, finality is the affirmation that all well-formed blocks will not be revoked once committed to the blockchain. For example, when 51% of the users agree on a false blockchain, they can change the valid structure.

Let us proceed to another potential weapon against fake news, artificial intelligence.

6. FAKE NEWS AND ARTIFICIAL INTELLIGENCE

Artificial intelligence algorithms in many cases can be used to distinguish algorithmically generated texts from human writings, but they are not prepared to verify the validity of information. If someone develops a machine learning tool for verification, the result will highly depend on the training datasets used for the learning process, and the bias introduced by the developers will affect the outcome (Council 2019).

When we discussed the potential application of blockchain technology to avoid fake information, we have not paid attention to the actual content of the news. Below we discuss the role of artificial intelligence, or more correctly, machine learning (ML) in connection with fake news. Machine learning is the second most powerful tool in both the generation and detecting of fake news, after human intelligence. But ML is much faster and less expensive than human intelligence.

In recent years ML methods have been developed to both create and detect false information. Up-to-date text generation algorithms are able to create realistic, but false research papers on almost any true or false fragments of input text which seem to be relevant for people who are not familiar with the particular topic area (SCIGen 2021).

ML methods for detecting false information use natural language analysis, based on feature extraction of collected corpora, but their efficiency is limited and they can be applied only by companies having the necessary knowledge and resources.

OpenAI, a research lab in 2019 released the full version of their text-generating AI system GPT-3, but experts warned it could be used for malicious purposes. Originally the full version of the system was withheld, "out of fear it would be used to spread fake news, spam, and disinformation" (Vincent 2019). As other similar intelligent algorithms were released, the company finally shared its product on the internet. GPT-3, an autoregressive natural language model, which contains 175 billion parameters, was tested on several datasets and demonstrated outstanding performance on many NLP tasks such as reading comprehension, question answering, textual entailment. Thus, GPT-3, beside its original functions, is a powerful tool for creating any fake texts provided the model is fine-tuned for the given topic (Brockmann et al. 2020). It was trained in an unsupervised way on millions of webpages from the WebText dataset.

Successful application of ML to generating or detecting fake news highly depends on the domain area. Let us consider the purpose of generating fake objects first, including deep fake, which means generation of fake images and videos. Fake news, images and videos are generated for three possible reasons: (1) scientific research; (2) business gain (e.g. click bait content[1] or marketing messages); (3) propaganda (political and business). We suppose that in scientific research fake objects and their generators are under full control and we will focus on the last two items. One can find alternative fact[2] and fake news generators on a number of websites, which are set up mostly for fun or educational purposes.

In the paper of Klein and Wueller (2017) fake news are defined "as the online publication of intentionally or knowingly false statements of fact." The question arises if a text produced by a natural language algorithm, such as GPT-3, which is not knowingly false, satisfies these criteria. Our answer is negative, GPT-3 model does not do anything intentionally or knowingly. But the output is in most cases fake.

Social sites have been developing machine learning algorithms for fake news detection. It is their business interest to minimize the amount of fake news e.g. conspiracy theories. Are they good enough? Below we refer to the most popular ML algorithms, developed for detecting fake news. Detection schemes may be grouped into two classes, linguistic-based methods and network-based methods (Conroy et al. 2015). Linguistic-based methods are mostly statistical methods and try to distinguish between the statistical properties of the valid and fake texts.

[1] Something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest (Merriam-Webster vocabulary).

[2] The term "alternative facts" was coined by U.S. Counselor to the President Kellyanne Conway during a Meet the Press interview on January 22, 2017.

Network-based methods use website information, author's data, etc. The two methods are frequently employed in combination.

Detecting fake news from a linguistic – natural language processing – view is a classification problem. A text is classified as valid or fake if we use a categorical classification, or the method provides a probability of belonging to one of the classes if the classification is probabilistic. For fake news detection the natural language processing concepts are applied from statistical analysis of Term-Document Matrices (TDMs; Antonellis – Gallopoulos 2006) to sophisticated neural networks. Each element of the TDM is the frequency of a given term in a given document and it is a suitable measure of the importance of each term with respect to the document and the entire collection. Effective methods for text mining and classification are for example latent semantic indexing (LSI; Papadimitriou et al. 2000) and singular value decomposition (SVD; Golub – Reinsch 1971). Traditional text mining approach will work only if the visible and latent factors of the TDM are different for valid and invalid information. This situation occurs only if the authors of fake news are not very sophisticated and the text is statistically different from valid news. In many cases commercial advertisements are statistically different from product specifications and can be successfully identified by LSI and SVD (Zhang 2014).

More complex machine learning models use analytical methods and network-based methods in combination. The models are overwhelmingly based on artificial neural networks and deep learning models where the features of the classified objects are typically latent, statistically not necessarily different from valid texts, and as the training datasets are collected from the web, it is difficult to categorize a method as linguistic or network-based.

Identification of originally unlabeled (i.e having no preliminary attached fake or valid labels) articles is a classification method, where, given a document, the system classifies it as fake or valid, based on the data collected from multiple sources. The routine training process of most machine learning methods requires preliminary human labeling of the training data.

But how can the analyst ensure that the human labeling is correct and the people doing the labeling can always properly differentiate between valid and fake training data? Hiriyannaiah et al. (2020), to avoid or reduce the bias in the GAN (Godfellow et al. 2014) method, proposed a method for labeling. GAN is constructed from two adversarial models, the Generator and the Discriminator. The Generator is a traditional ML model, which uses an unlabeled real dataset of valid and maybe fake news, and generates synthetic texts, while the Discriminator evaluates the output for validity and gives feedback to the Generator if the output is valid or fake. Both Generator and Discriminator are neural networks. During the iteration of generating and evaluating, the Generator's output the Discriminator's ability to differentiate between fake and valid news is also improved. Goodfellow et al. demonstrated that the GAN model can generate images that seem original to human viewers (Jones 2017). A GAN created painting called "Edmond de Belamy," was sold for $432,500.

In the text classification method, both the Generator and the Discriminator learn from the trial. This step is repeated several times. The goal of the Generator is to generate an output which cannot be accurately classified by the Discriminator; i.e the Discriminator only can predict the class (true or false) with equal probabilities, which means that the generated text might be equally true or false.

The Generator agent maps the data points (feasible texts) with latent features as their coordinates to the target space, and the mapping is consequently evaluated by the Discriminator. The loss function, which is the measure of accuracy of the algorithm in the training process consists of two parts, the first part is the loss function of the Discriminator (e.g. expected log-likelihood of the sample) which is maximized by the Discriminator, and the second part is the expected log-likelihood of the generated sample. The overall loss function is the sum of the two parts. The Discriminator's goal is to maximize the overall loss, while the goal of the Generator is to minimize it:

$$\min_{G} \max_{D} l(G, D)$$

Originally, the GAN model was developed for continuous models and successfully applied to image generations, as image features (e.g.) are continuous. Text datasets however are not continuous. To solve the problem of discrete values, the SeqGAN method was developed in 2017 (Yu et al. 2017). A second problem of the Discriminator that it can only evaluate a finished text proposed by the Generator. The sequence generator (SeqGAN) however, generates stochastic steps in a reinforced learning framework (Sutton – Barto 1998). These steps are the gradient policy upgrades, where the Reinforcement Learning reward values come from the Discriminator. A policy is defined as the probability distribution of actions in a given state. The objective of a Reinforcement Learning agent is to maximize the "expected" reward when following a policy, depending on some parameters. The optimization is performed in ML using gradient descent.

The resulting classifier identifies data as fake or valid. The classifier is a complex neural network with high accuracy. It is trained to identify the synthetic texts generated by the Generator and discriminate it from the real texts of the dataset. The trained model is used for prediction. The classifier is then trained to identify the real examples and the examples generated by the Generator.

The principal problem with the application of SeqGAN for fake news detection is that we only know that the data in the training are real but not necessarily valid. But the overall outcome heavily depends on the fact if the data objects are valid or fake. The GAN method will discriminate with high accuracy between real and synthetic texts, but real texts on the input side could also be fake and the method will not be effective.

7. CONTENT ANALYSIS

Most direct content analysis schemes are customized for special data types, thereby they are adequate for content-specific fake news detection systems (Zhang et al. 2019).

A rumor detection model (Qazvinian et al. 2011) uses various features of the content, e.g. words, parts of speech, segmentation) and network-specific properties, as hashtags, URLs, retweets on Twitter, etc. This approach is also employed for the identification of fake information in microblogs.

Rubin et al. (2016) devised an SVM-based algorithm, AHGNA, that embraces five predictive features (Absurdity, Humor, Grammar, Negative Affect, and Punctuation) and trained it on 360 news articles with interesting results in specific cases.

To solve the problem of generality, Zhang et al. (2019) propose creating clusters of trustworthy news according to topics, and a new object is identified as fake if it is not a part of any topic cluster or the similarity between the relevant topic clusters and the new object is below a predefined threshold. The cluster members are preliminary labeled as valid by the research community or they come from valid sources.

And thus we have returned to the Byzantine General's Problem. What if the validators behave as the Byzantine Generals? And, who will select, and how, the trustworthy sources? As we have seen in the Covid-19 communication, several "official" sources issued contradictory information, sometimes contradicting themselves in short timeframes.

On the opinion-neutral media platforms, people tend to believe others with similar viewpoints, in this case the sufficient conditions for the solution of the Byzantine Generals' Problem is met only if more than three quarters of the researchers have identical views, which are valid. Unfortunately, the frequent fact-checking actions of the media might have little effect on people. They are ineffective in reducing false information consumption (Wood – Porter 2019).

Other methods of detecting fake news are related to statistical analysis of speech (Hancock et al. 2013) which identifies the speech of specific groups. Rhetorical structure theory (Rubin – Lukoianova 2015) or RST is involved to identify the difference between valid and fake texts with the use of the Vector Space Model (Venkat – Amogh 2018)

9. LEARNING FROM POSITIVE AND UNLABELED EXAMPLES (PU LEARNING)

PU learning is a learning model which is trained on positive and unlabeled data. Unlabeled data might be either positive or negative. It builds a binary classifier that classifies the test data into two classes, positive and negative (Fusilier et al. 2015). The PU learning is successfully applied to spam filtering. The bottleneck is the human labeling of the positive data. How to reach agreement on labeling if the texts are not simple spam, but sophisticated fake news?

Linguistic cue analysis (Siering et al. 2016) is another content-based approach to fake news detection, which analyze the content of deceptive content on crowd-funding projects, focusing on the differences of static and dynamic communications.

There are several other methods based on content analysis, which are very useful in specific topic areas, however they all require human labeling, which we try to avoid in order to prevent bias.

One can ask how to apply the sophisticated NLP models for low-resource languages, i.e. languages spoken in smaller regions which cannot invest enough into NLP research and thus lack datasets for the training of the ML models. Low-resource languages have disadvantage in natural language processing, because the vast majority of intellectual and material resources in the research of natural language processing are concentrated around high-resource languages, English, Spanish, Chinese, etc. But the main theoretical models and free softwares are freely available for each researcher and easily applicable to small languages. The best language models, however, such as BERT or GPT3 are originally trained in English. They require large amount of training data and/or sophisticated language-specific engineering. Such amount of data is unavailable for the smaller languages and, in many cases, you cannot find a linguistically trained speaker to build a language model.

There are two main approaches to use effective NLP in the low-resource setting, where the amount of data and the knowledge of the language are insufficient for traditional approaches. These are either traditional approach that focuses on collecting more data for a single language or a variety of languages, or approaches that apply transfer learning (SCIFORCE 2018).

The traditional approach is to build large corpora for each particular language. The drawback of this approach is that language-specific corpora are not transferable to other languages. More general approaches are the Human Language Project, (Abney – Bird 2010) which aims at standardizing the formats of the corpora and the Leipzig Corpora Collection (Biemann et al. 2020) which includes corpora for 290 languages.

10. SERVICES FOR IDENTIFYING FAKE NEWS

There are some freely available applications and services helping to discover fake news for the users. These methods are not automatic, as they use datasets for training or reference selected and labeled by humans.

- *B.S. Detector,* a Chrome extension that highlights fake news links. Data on fake domains comes from a list of fake news sources, such as https://github.com/lishiyo/bs-detector, but there are many similar sources available.
- *PolitiFact,* is a website (http://politifacts.com) collecting information from journalists, politicians and simple readers to check facts. The site always displays the source documents of the valid information. The PolitiFact system largely depends on human intervention.
- The search engine approach used by Google is based on the page ranking algorithm, which measures the relevance of a webpage. The relevance is based on the number of external links. One can consider a webpage with high relevance as valid, while with low relevance as fake. But fake news in many cases can have more attention and external links than valid ones.
- Google also developed a technology that scores the accuracy of facts presented by web pages. The technology tries to categorize the content in context automatically without human labeling
- *Weigh Facts* is an NLP algorithm that processes the text and compares the elements with other webpages. If the similarity is high with reputed sources, the text receives a high score.
- *Reputation Prediction* predicts the reputation of a webpage by a ML algorithm from a set of features, including web ranks.
- Use of *Sensational Words* is based on how a statistical analysis of keywords in the headlines sometimes identifies the fake news with significant success.
- *Crosscheck* was a French fact-checking service to identify fake news during the last French election, and it evolved to another fact-checking project FirstDraw. Its mission is to protect communities from harmful misinformation in a collaboration with a global network of journalists who investigate and verify emerging stories.

Further AI based tools and services include the following.

- *Spike* monitors Facebook, Twitter, Instagram, Pinterest, YouTube, and the web in real time and predicts breakout and viral stories. Spike does not detect fake news, it just predicts the mainstream news.
- *Hoaxy* is a collaborative tool that tries to identify sites with fake news. It visualizes the spread of claims and fact checking by drawing the spread of the news on the internet and Twitter. It colour codes human-like communication and robot-like communication.
- *Snopes* is a fact-checking site. The hot topic news are categorized into real and fake by human researchers.

- *CrowdTangle* is a tool from Facebook to help follow, analyze, and report on what is happening across social media. In 2019, Facebook, Instagram and Reddit started a pilot project in partnership with researchers and academics to help study the spread of public content on Facebook and Instagram. Researchers use CrowdTangle to analyze critical topics such as racial justice, misinformation and abuse of social platforms.
- *Check* helps to verify online breaking news. The Meedan company started Check, a global fact-checking project in 2019 using the WhatsApp Business API in 5 countries and 4 languages with the support of Facebook and Whatsup.
- *Le Decodex* from Le Monde is a database of fake and real websites.
- *Pheme* is an EU project to develop the technology to check the veracity of user-generated and online content. Phame was the Greek goddess of fame and rumours. The main objective of the project is to develop technology to identify phemes (rumorous memes) and model their spread over social and online media. The types to be identified are speculation, controversy, misinformation, and disinformation. It is very hard to identify a pheme and interpret its context automatically. The Pheme project uses an interdisciplinary approach, including NLP, text mining, web science, analysis of social networks, and visualization. The lexical, semantic and syntactic information in the document are analyzed, then the elements are compared with trustworthy data sources, such as PubMed, Linked Open Data, Ontotext, and GraphDB. The diffusion of the information is also analyzed to check reliability. The project aims at releasing open source veracity intelligence algorithms, complemented by a human-analyzed rumour dataset. This will be used as “rumor intelligence”, that is the ability to identify rumours in near real time. The projected direct application areas are medical information systems and digital journalism.

We can predict the very fast growth of AI’s success in detecting fake news created by humans, but perhaps, detecting synthetic news created by contemporary NLP algorithms is much harder for an AI agent. AI-enabled computational disinformation has been used for a few years now. Hackers created effective AI bots over social media. Social bots are not illegal, some websites even support using them, e.g. Facebook has an Official API for writing bots and if the content satisfies the Terms of Service, the site will not delete it. But hackers can design phishing campaigns, propaganda, etc. It is the responsibility of social sites to set up limitations.

The question is whether AI is or will be a perfect tool for detecting fake news. In our opinion it is a very good tool, but today it is far from perfect and perhaps never will be perfect. Let us summarize the benefits and deficiencies:

Benefits of the application of a contemporary AI algorithm for detecting fake news:

- Very large labeled datasets are available for training and verifying the models.
- Some models have billions of parameters and reach very high accuracy in the areas where they were trained.
- Researchers can use the popular transfer learning method to build accurate models. The transfer learning is very popular in NLP (Google Bert and Transformer-XL, Stanford NLP, Open-AI’s GPT-3, etc.) and image processing (VGG-16, ResNet50, Inceptionv3, etc.) The pre-trained models are trained on a large dataset with large computers and one can import these models in their own program, saving time and expenses.

Deficiencies of the application of AI:

- The AI classification models are not general. If a fake news detection program was trained on political data, it will not be accurate on health news. Even the pre-trained models with billions of parameters are relevant only to topic areas where the training datasets were collected.
- The training datasets – no matter how large – are collected by humans and might be biased. If the dataset is collected automatically by automatic web scraping, it is also biased by the fake datasets which are always present on the net.
- Labeling of the training data is decisive for the outcome. Labels in most cases are attached by people, and they can make mistakes. When AI algorithms are used for labeling, the initial labeled input has a determining role.
- The algorithms are taught a sense of values of the teachers (Booch 2016).

Can we use AI to fight fake news at all? The answer is a definitive yes, but the quality of the output depends on a few factors:

- Has the AI model been developed for the specific topic area?
- Is the initial training dataset unbiased, e.g. approved by trustworthy fact-checking?
- Is the model sophisticated enough to classify texts from fake sources which are familiar to those the AI algorithms used for detecting fake news? Social sites are continuously developing their fake news detecting algorithms, but fake news sources also develop their own methods, to deceive the detecting algorithms. The temporary winner is the one who has more knowledge and resources. But there is no final victory, the competition goes on. And, who will decide for a social site, what is fake and what is not? The site itself, and one can only hope that they make a correct decision.

11. CONCLUSION

There is no perfect weapon against fake news. Statistical analysis of the text and statistical classification methods are only useful for filtering out the majority of not very sophisticated attempts for falsifications.

Network methods with automatic cross-checking of texts with other publications are an effective method if the Byzantine General's conditions are met. However, if the quantity of the fake news exceeds a threshold, the cross checking loses its power. Excellent visualization tools are available to follow the spread of a news statement typed in by the user of the service. In most cases the unique source will be found in real time. Knowing the source, the user can decide if the source is reliable or not.

The AI algorithms are very promising for real time classification of news, but the input training datasets are created by people and they have to be verified, which is a very expensive task. And at the end, the classification will be driven by the opinion of the authors of the training datasets.

Blockchain technology is very good for verification of news in a well-defined and organized community which shares similar views. In this community, the application of a proper consensus protocol proves with a very high probability for the participants – and for them only – that the text, approved by a previously nominated and trusted body, was not changed by anyone intentionally or unintentionally.

Fact-checking services are effective for journalists, bloggers, etc. They present the original sources of a fact under checking, as written documents, videos and audio files. The question again, how do we know if the sources themselves are not fake. If a fact-checking service has a trustworthy reputation, we believe in the sources they present.

The successful detection of fake news perhaps is a combination of several methods, each having its function in the verification process.

The role of social sites in spreading and also detecting fake news is essential, as for many people social sites are the basic source of information. They can employ statistical analysis, network method, machine learning and fact checking in combination. There is another problem with social sites as exclusive sources of information. There is a debate over the proposed action if fake news is detected by a social site or other publishers. A part of the users would like to have the fake news eliminated while others prefer passive identification.

It is also critical that online media checks are conducted after the publications when the false text has already gone viral. The large social sites spend a lot of resources, including human moderation, to discover and take down toxic speech, such as terrorist content, but they do spend less on the discovery and taking down or, at least marking, of fake content.

Suppose we have successfully identified a piece of fake news. How to proceed? Identification of fake information is partly a scientific problem, but deleting them from the internet is a philosophical, business and political issue. What is more important, freedom of speech, or not allowing people to broadcast theories about microchips administered to our body through the Covid vaccine? This could be the topic of a very interesting discussion, but it is beyond the framework of this paper.

ACKNOWLEDGEMENT

The research was supported by GINOP-2.2.1-18-2018-0010.

REFERENCES

Abney, S. – Bird, S. (2010): The Human Language Project. Building a Universal Corpus of the World’s Languages. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 88-97.

Altay, S. – Hacquin, A. S. – Mercier, H. (2020): Why Do So Few People Share Fake News? It Hurts Their Reputation. *New Media & Society* https://doi.org/10.1177/1461444820969893.

Antonellis, I. – Gallopoulos, E. (2006): Exploring Term-Document Matrices from Matrix Models in Text Mining. http://archive.siam.org/meetings/sdm06/workproceed/Text%20Mining/antonellis21.pdf, accessed 04/10/2021.

Bashir, I. (2020): *Mastering Blockchain*. Packt Publishing Co.

Biemann, C. – Heyer, G. – Quasthoff, U. – Richter, M. (2020): The Leipzig Corpora Collection: Monolingual Corpora of Standard Size. https://corpora.uni-leipzig.de/de?corpusId=deu_news_2020, accessed 04/10/2021.

Booch G. (2016): Don’t Fear Superintelligent AI. www.ted.com/talks/grady_booch_don_t_fear_superintelligent_ai?language=en, accessed 04/10/2021.

Brockman, G. – Murati, M. – Welinder, P. (2020): OpenAI API. https://openai.com/blog/openai-api/, accessed 04/10/2021.

Burkhardt, J. M. (2017): Combating Fake News in the Digital Age. *Library Technology Reports November/December*.

Cappacioli, G. (2019): Blockchain: Proof of Authority (PoA). https://affidaty.io/blog/en/2019/08/blockchain-proof-of-authority-poa/#:∼:text=There%20is%20a%20different%20type,PoW%20system%20of%20Ethereum%20itself, accessed 14/10/2021.

Castro, M. – Liskov, B. (1999): Practical Byzantine Fault Tolerance. Proceedings of the Third Symposium on Operating Systems Design and Implementation, New Orleans, USA.

Chen, L. – Xu, L. – Shah, N. – Gao, Z. – Lu, Y. – Shi, W. (2017): On Security Analysis of Proof-of-Elapsed-Time (PoET). International Symposium on Stabilization, Safety, and Security of Distributed Systems.

Colemen, A. (2020): Hundreds Dead Because of Covid-19 Misinformation. https://www.bbc.co.uk/news/world-53755067, accessed 04/10/2021.

Conroy, N. J. – Rubin, V. L. – Chen, Y. (2015): Automatic Deception Detection: Methods for Finding Fake News. *Proceedings of the Association for Information Science and Technology* 52(1): 1–4.

Council G. (2019): The Machine Learning Data Dilemma. https://tdwi.org/articles/2019/04/15/adv-all-machine-learning-data-dilemma.aspx, accessed 04/10/2021.

Del Vicario, M. – Bessi, A. – Zollo, F. (2016): The Spreading of Misinformation Online. *Proceedings of the National Academy of Sciences* 113(3): 554–559.

Diamantopoulos, P. – Maneas, S. – Patsonakis, C. – Roussopoulos, M. (2015): Interactive Consistency in Practical, Mostly-Asynchronous Systems. 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS).

Dwork, C. – Naor, M. (1992): Advances in Cryptology. In: *Lecture Notes in Computer Science* 740. Berlin: Springer.

Dziembowski, S. – Faust, S. – Kolmogorov, V. – Pietrzak, K. (2013): Proofs of Space. http://eprint.iacr.org/2013/796, accessed 04/10/2021.

Fusilier, D. H. – Montes-y Gómez, M. – Rosso, P. – Cabrera, R. G. (2015): Detecting Positive and Negative Deceptive Opinions Using Pu-Learning. *Information Processing & Management* 51(4): 433–443.

Golub, G. H. – Reinsch, C. (1971): Singular Value Decomposition and Least Squares Solutions. In: Bauer, F. L. (ed.): *Linear Algebra. Handbook for Automatic Computation*, vol. 2. Berlin: Springer.

Goodfellow, I. – Pouget-Abadie, J. – Mirza, M. – Xu, B. – Warde-Farley, D. – Ozair, S. – Courville, A. – Bengio, Y. (2014): Generative Adversarial Networks. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*, pp. 2672–2680.

Hayes, A. (2021): Proof of Capacity https://www.investopedia.com/terms/p/proof-capacity-cryptocurrency.asp, accessed 04/10/2021.

Hancock, J. T. – Woodworth, M. T. – Porter, S. (2013): Hungry Like the Wolf: A Word–Pattern Analysis of The Language of Psychopaths. *Legal and Criminological Psychology* 18(1): 102–114.

Harari, Y. N. (2018): Are We Living in a Post-Truth Era? Yes, But that's because We're a Post-Truth Species. https://ideas.ted.com/are-we-living-in-a-post-truth-era-yes-but-thats-because-were-a-post-truth-species/, accessed 04/10/2021.

Hiriyannaiah, S. – Srinivas, A. M. D. – Shetty, G. K. – Siddesh, G.M. – Srinivasa, K. G. (2020): A Computationally Intelligent Agent for Detecting Fake News Using Generative Adversarial Networks. In: Bhattacharyya, S. – Snášel, V. – Gupta, D. – Khanna, A. (eds): *Hybrid Computational Intelligence for Pattern Analysis and Understanding.* Academic Press, pp. 69-96.

Islam, S. – Sarkar, T. – Khan, S. H. – Kamal, M. – Hasan, M. – Kabir, A. – Yeasmin, D. –Islam, M. A. – Chowdhury, K. – Anwar, K. S. – Chughtai, A. A. – Seale, H. (2020): COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *Amercian Journal of Tropical Medicince and Hygene* 103(4): 1621–1629.

Jones, K. (2017): Creating Art with GANs (2017). https://towardsdatascience.com/gangogh-creating-art-with-gans-8d087d8f74a1, accessed 04/10/2021.

Karantias, K. – Kiayias, A. – Zindros, D. (2017): Proof-of-Burn. https://eprint.iacr.org/2019/1096.pdf, accessed 04/10/2021.

Klein, D. O. – Wueller, J. R. (2017): Fake News: A Legal Perspective. *Journal of Internet Law* 20(10): 1–13.

Lamport, L. – Shostak, R. – Pease, M. (1982): The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems* 4(3): 382–401.

Landemore, H. (2012): Collective Wisdom Old and New. In: Landemore, H. – Elster, J. (eds): *Collective Wisdom. Principles and Mechanisms*. Cambridge: Cambridge University Press.

Lewandowsky, S. – Ecker, U. K. – Cook, J. (2017): Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era. *Journal of Applied Research in Memory and Cognition* 6(4): 353–369.

Malkhi, D. – Reiter, M. (1998): Byzantine Quorum Systems. *Distributed Computing* 11: 203–213.

MacDonald, E. (2017): The Fake News that Sealed the Fate of Antony and Cleopatra https://theconversation.com/the-fake-news-that-sealed-the-fate-of-antony-and-cleopatra-71287, accessed 04/10/2021.

Merkle, R. C. (1982): *Method of Providing Digital Signatures*, US patent 4309569.

Merkle, R. C. (1988): A Digital Signature Based on a Conventional Encryption Function. Advances in Cryptology — CRYPTO ’87. In: *Lecture Notes in Computer Science* 293. Berlin: Springer.

Mosseri, A. (2017): Working to Stop Misinformation and False News. Facebook for Media https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news, accessed 04/10/2021.

Papadimitriou, C. H. – Raghavan, P. – Tamaki, H. – Vempala, S. (2000): Latent Semantic Indexing. *Journal of Computer and System Sciences* 61: 217–235.

Pariser, E. (2012): *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin Books.

Qazvinian, V. – Rosengren, E. – Radev, D. R. – Mei, Q. (2011): Rumor has it: Identifying Misinformation in Microblogs. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1589–1599.

Racsko, P. (2019): Blockchain and Democracy. *Society and Economy* 41(3): 353–369.

Reiff, N. (2021): What’s the Environmental Impact of Cryptocurrency? https://www.investopedia.com/tech/whats-environmental-impact-cryptocurrency/, accessed 04/10/2021.

Rubin, V. L. – Lukoianova, T. (2015): Truth and Deception at the Rhetorical Structure Level. *Journal of the Association for Information Science and Technology* 66(5): 905–917.

Rubin, V. – Conroy, N. – Chen, Y. – Cornwell, S. (2016): Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pp. 7–17.

SCIFORCE (2018): Nlp for Low Resource Settings: https://medium.com/sciforce/nlp-for-low-resource-settings-52e199779a79, accessed 04/10/2021.

SCIgen (2021): SCIgen - An Automatic CS Paper Generator. https://pdos.csail.mit.edu/archive/scigen/, accessed 04/10/2021.

Shargel R. – Twiss, L. (2019): Evidenced-Based Thinking for Scientific Thinking. In: Murtonen, M. – Balloo, K. (eds): *Redefining Scientific Thinking for Higher Education*. Cham: Palgrave Macmillan.

Shi, P. – Wang, H. – Yang, S. – Chen, C. – Yang, W. (2021): Blockchain-Based Trusted Data Sharing Among Trusted Stakeholders in IoT. *Software: Practice and Experience* 51: 2051– 2064.

Siering, M. – Koch, J.-A. – Deokar, A. V. (2016): Detecting Fraudulent Behavior on Crowdfunding Platforms: The Role of Linguistic and Content-Based Cues in Static and Dynamic Contexts. *Journal of Management Information Systems* 33(2): 421–455.

Susarla, A. (2021): Big Tech has a Vaccine Misinformation Problem – Here's What a Social Media Expert Recommends. https://theconversation.com/big-tech-has-a-vaccine-misinformation-problem-heres-what-a-social-media-expert-recommends-164987, accessed 04/10/2021.

Sutton, R. S. – Barto, A. G. (1998): *Reinforcement Learning: An Introduction*. Boston: MIT Press.

VeChain Foundation (2019): VeChain Whitepaper 2.0. http://www.vechain.org/qfy-content/uploads/2020/01/VeChainWhitepaper_2.0_en.pdf, accessed 04/10/2021.

Vedder, A. – Wachbroit, R. (2003): Reliability of Information on the Internet: Some Distinctions. *Ethics and Information Technology* 5: 211–215.

Venkat, N. G. – Amogh, R. G. (2018): Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications. In: Handbook of Statistics, 1st ed. 38. Elsevier.

Vincent, J. (2019): OpenAI has Published the Text-Generating AI it Said was too Dangerous to Share. https://www.theverge.com/2019/11/7/20953040/openai-text-generation-ai-gpt-2-full-model-release-1-5b-parameters, accessed 04/10/2021.

Watson, C. A. (2018): Information Literacy in a Fake/False News World: An Overview of the Characteristics of Fake News and its Historical Development. *International Journal of Legal Information* 46(2): 93–96.

Wendling, M. (2018): *The (Almost) Complete History of 'Fake News.'* BBC News https://www.bbc.co.uk/news/blogs-trending-42724320, accessed: 04/10/2021.

Wenting, L. - Andreina, S. – Bohli, J. M. – Karame, G. (2017): Securing Proof-of-Stake Blockchain Protocols. In Garcia-Alfaro, J. – Navarro-Arribas, G. – Hartenstein, H. – Herrera-Joancomartí, J. (eds.): *Data Privacy Management, Cryptocurrencies and Blockchain Technology. Lecture Notes in Computer Science*. Cham: Springer, pp. 297–315.

Whitty, M. T. – Joinson, A. (2008): *Truth, Lies and Trust on the Internet*. London: Routledge.

Wood, T. – Porter, E. (2019): The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior* 41: 135–163.

Yu L. – Zhang, W. – Wang, J. – Yu, Y. (2017): *Sequence Generative Adversarial Nets with Policy Gradient*. https://arxiv.org/abs/1609.05473, access 04/10/2021.

Zhang, C. – Gupta, A. – Kauten, C. (2019): Detecting Fake News for Reducing Misinformation Risks Using Analytics Approaches. *European Journal of Operational Research* 279: 1036–1052.

Zhang J. – Rong Y. – Cao, J. – Rong, C. – Bian J. – Wu ,W. (2021): DBFT: A Byzantine Fault Tolerance Protocol with Graceful Performance Degradation. *IEEE Transactions on Dependable and Secure Computing*, https://doi.org/10.1109/TDSC.2021.3095544.

Zhang, D. (2014): Detecting Ads in a Machine Learning Approach. http://cs229.stanford.edu/proj2014/Di%20Zhang,%20Detecting%20Ads%20in%20a%20Machine%20Learning%20Approach.pdf, accessed 04/10/2021.

Zimwara, T. (2020): $5.6 Million Double Spent: ETC Team Finally Acknowledges the 51% Attack on Network. https://news.bitcoin.com/5-6-million-stolen-as-etc-team-finally-acknowledge-the-51-attack-on-network/, accessed 04/10/2021.