Advanced Machine Learning Techniques for Fake News (Online Disinformation) Detection: A Systematic Mapping Study

Michał Choraś[a], Konstantinos Demestichas[b], Agata Giełczyk[a], Álvaro Herrero[c], Paweł Ksieniewicz[d], Konstantina Remoundou[b], Daniel Urda[c], Michał Woźniak[d,*]

[a]*UTP University of Science and Technology, Poland*
[b]*National Technical University of Athens, Greece*
[c]*Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad de Burgos, Av. Cantabria s/n, 09006, Burgos, Spain.*
[d]*Wrocław University of Science and Technology, Poland*

Abstract

Fake news has now grown into a big problem for societies and also a major challenge for people fighting disinformation. This phenomenon plagues democratic elections, reputations of individual persons or organizations, and has negatively impacted citizens, (e.g., during the COVID-19 pandemic in the US or Brazil). Hence, developing effective tools to fight this phenomenon by employing advanced Machine Learning (ML) methods poses a significant challenge. The following paper displays the present body of knowledge on the application of such intelligent tools in the fight against disinformation. It starts by showing the historical perspective and the current role of fake news in the information war. Proposed solutions based solely on the work of experts are analysed and the most important directions of the application of intelligent systems in the detection of misinformation sources are pointed out. Additionally, the paper presents some useful resources (mainly datasets useful when assessing ML solutions for fake news detection) and provides a short overview of the most important RD projects related to this subject. The main purpose of this work is to analyse the current state of knowledge in detecting fake news; on the one hand to show possible solutions, and on the other hand to identify the main challenges and methodological gaps to motivate future research.

Keywords: Fake news, Machine Learning, Social media, Media content manipulation, Disinformation detection

1. Introduction

Let us start with a strong statement: the fake news phenomenon is currently a big problem for societies, nations and individual citizens. Fake news has already plagued democratic elections, reputations of individual persons or organizations, and has negatively impacted citizens in the COVID-19 pandemic (e.g., fake news on alleged medicines in the US or in Brazil). It is clear we need agile and reliable solutions to fight and counter the fake news problem. Therefore, this article demonstrates a critical scrutiny of the present level of knowledge in fake news detection, on one hand to show possible solutions but also to motivate the future research in this domain.

Fake news is a tough challenge to overcome, however there are some efforts from the Machine Learning (ML) community to stand up to this harmful phenomenon. In this mapping study, we present such efforts, solutions and ideas. As it is presented in Fig. 1, fake news detection may be performed by analysing several types of digital content such as images, text and network data, as well as the author/source reputation.

This survey is not the first one in the domain of fake news. Another major comprehensive work addressing the ways to approach fake news detection (mainly text analysis-based) and mainstream fake news datasets is [1].

*Corresponding author
Email addresses: chorasm@utp.edu.pl (Michał Choraś), cdemest@cn.ntua.gr (Konstantinos Demestichas), agata.gielczyk@utp.edu.pl (Agata Giełczyk), ahcosio@ubu.es (Álvaro Herrero), pawel.ksieniewicz@pwr.edu.pl (Paweł Ksieniewicz), kremoundou@cn.ntua.gr (Konstantina Remoundou), durda@ubu.es (Daniel Urda), michal.wozniak@pwr.edu.pl (Michał Woźniak)

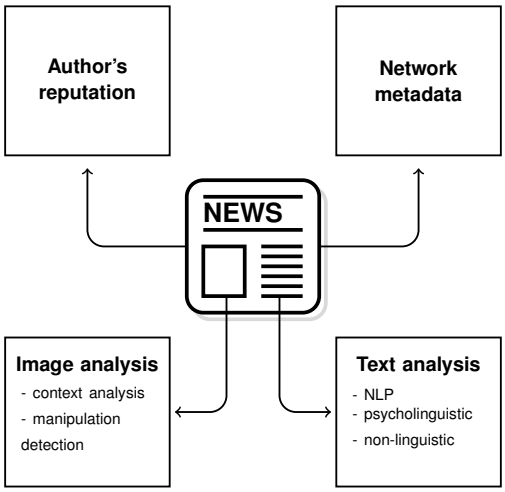

Figure 1: The types of digital content that are analysed so as to detect fake news in an automatic manner

According to it, the *state-of-the-art* approaches for this kind of analysis may be classified into five general groups with methods relying upon: (*i*) linguistic features, (*ii*) deception modelling, (*iii*) clustering, (*iv*) predictive modelling and (*v*) content cues. With regard to the text characteristics, style-based and pattern-based detection methods are also presented in [2]. Those methods rely on the analysis of specific language attributes and the language structure. The analyzed attributes found by the authors of the survey include such features as: quantity of the language elements (e.g. verbs, nouns, sentences, paragraphs), statistical assessment of language complexity, uncertainty (e.g. number of quantifiers, generalizations, question marks in the text), subjectivity, non-immediacy (such as the count of rhetorical questions or passive voice), sentiment, diversity, informality and specificity of the analyzed text. Paper [3] surveys several approaches to assessing fake news, which stem from two primary groups: linguistic cue approaches (applying ML) as well as network analysis approaches.

Yet another category of solutions is network-based analysis. In [4], two distinct categories are mentioned: (*i*) social network behavior analysis to authenticate the news publisher's social media identity and to verify their trustworthiness and (*ii*) scalable computational fact-checking methods based on knowledge networks. Beside text-based and network-based analysis, some other approaches are reviewed. For example, [5] attempts to survey identification and mitigation techniques in combating fake news and discusses feedback-based identification approaches.

Crowd-signal based methods are also reported in [6], while content propagation modelling for fake news detection purposes, alongside credibility assessment methods, are discussed in [2]. Such credibility-based approaches are categorized here into four groups: evaluation of news headlines, news source, news comments and news spreaders/republishers. In addition, in some surveys, content-based approaches using non-text analysis are discussed. The most common ones are based on image analysis [1, 5].

As complementary to the mentioned surveys, the present paper is unique by catching a very different angle of fake news detection methods (focused on advanced ML approaches). Moreover, in addition to overviewing current methods, we propose our own analysis criterion and categorization. We also suggest expanding the context of methods applicable for such a task and describe the datasets, initiatives and current projects, as well as the future challenges.

The remainder of the paper is structured in the following manner: in Section 1, previous surveys are overviewed and the historic evolution of fake news is presented, its current impact as well as the problem with definitions. In Section 2, we present current activities to address the fake news detection problem as well as technological and educational actions. Section 3 constitutes the in-depth systematic mapping of ML based fake news detection methods focused on the analysis of text, images, network data and reputation. Section 4 describes the relevant datasets used nowadays. In the final part of the paper we present some most emerging challenges in the discussed domain and we draw the main conclusions.

1.1. A historic perspective

Even though the fake news problem has lately become increasingly important, it is not a recent phenomenon. According to different experts [7], its origins are in ancient times. The oldest recorded case of spreading lies to gain some advantage is the disinformation campaign that took place on the eve of the *Battle of Kadesh*, dated around 1280 B.C., where the Hittite Bedouins deliberately got arrested by the Egyptians in order to tell *Pharaoh Ramses II* the wrong location of the *Muwatallis II* army [8].

Long time after that, in 1493, the Gutenberg printing press was invented. This event is widely acknowledged as a keystone in the history of news and press media, as it revolutionized this field. As a side effect, the dis- and misinformation campaigns had immeasurably intensified. As an example, it is worth mentioning the *Great Moon Hoax*, dating back to 1835. This term is a reference to the collection of half a dozen papers published in *The Sun*, the newspaper from New York. These articles concerned the ways life and culture had allegedly been found on the Moon.

More recently, fake news and disinformation played a crucial role in World War I and II. On the one hand, British propaganda during World War I was aimed at demonising German enemies, accusing them of using the remains of their troops to obtain bone meal and fats, and then feeding the rest to swines. As a negative consequence of that, Nazi atrocities during World War II were initially doubted [9].

On the other hand, fake news was also generated by Nazis for sharing propaganda. Joseph Goebbels, who co-operated closely with Hitler and was responsible for German Reich's propaganda, performed a deciding role in the news media of Germany. He ordered the publication of a paper *The Attack*, which was then used to disseminate brainwashing information. By means of untruth and misinformation, the opinion of the public was being persuaded to be in favour of the dreadful actions of the Nazis. Furthermore, according to [10], to this day, it has been the most disreputable propaganda campaign ever mounted.

Since the Internet and the social media that come with it became massively popularized, fake news has disseminated at an unprecedented scale. It is increasingly impacting on presidential elections, celebrities, climate crisis, healthcare and many other topics. This raising popularity of fake news may be easily observed in Fig.2. It presents the count of records in the *Google Scholar* database appearing year by year (since 2004), related to the term "fake news".

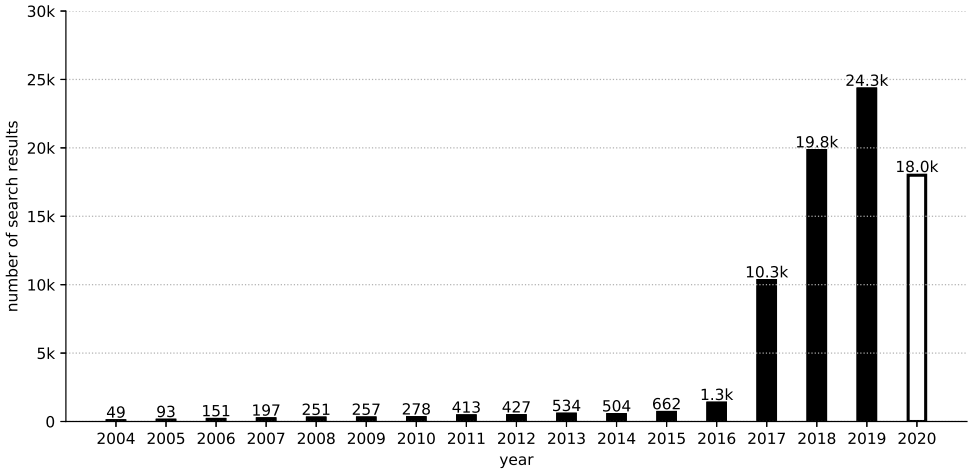

Figure 2: Evolution of the number of publications *per year* retrieved from the keyword "*fake news*" according to *Google Scholar*. For the year 2020, the status as of September, 8th.

1.2. Overview of definitions: what is meant by fake news?

Defining what the fake news really is poses a significant challenge. As a starting point, it is worth mentioning that Olga Tokarczuk, during her 2018 Nobel lecture [1], said:

[1] `https://www.nobelprize.org/prizes/literature/2018/tokarczuk/104871-lecture-english/`

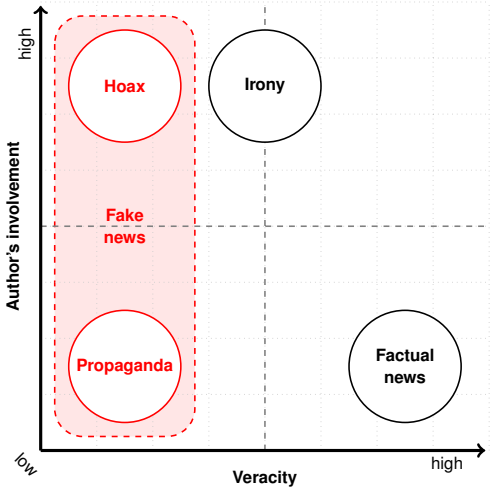

Figure 3: Fake news in the context of information

> *"Information can be overwhelming, and its complexity and ambiguity give rise to all sorts of defense mechanisms—from denial to repression, even to escape into the simple principles of simplifying, ideological, party-line thinking. The category of fake news raises new questions about what fiction is. Readers who have been repeatedly deceived, misinformed or misled have begun to slowly acquire a specific neurotic idiosyncrasy."*

There are many definitions of fake news [11]. In [12], it is defined as follows: *'the news articles that are intentionally and verifiably false, and could mislead readers'*. Quoting *Wikipedia*, being quite more verbal and less precise, it is: *'a type of yellow journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media'*. On the other hand, in Europe, the European Union Agency for Cybersecurity (*ENISA*) is using the term '*online disinformation*' when talking about fake news[2]. The *European Commission*, on their websites, describes the fake news problem as *'verifiably false or misleading information created, presented and disseminated for economic gain or to intentionally deceive the public'* [3]. This lack of a standardized and widely accepted interpretation of the fake news term has been already mentioned in some previous papers [13]. It is important to remember that conspiracy theories are not always considered as fake news. In such cases, the text or images found in the considered piece of information have to be taken into account along with the motivation of the author/source.

In classification tasks it is very important to distinguish between deliberate deception (actual fake news) and irony or satire that are close to it, but completely different in the author's intention. The difference is so blurred that it is sometimes difficult even for people (especially those without a specific sense of humor), so it is a particular problem for automatic recognition systems. The definition is therefore difficult to establish, but indeed fake news is a concept related to information; therefore, we tried to position it within some other information concepts, as presented in Fig. 3.

Factual news is based on facts concerned with actual details or information rather than ideas or feelings about it.

1.3. Why is fake news dangerous?

During the pandemic of Coronavirus (COVID-19) in 2020 we have had the opportunity to experience the disinformation power of fake news in all its infamous glory. The *World Health Organization* called this side-phenomena the '*infodemic*' – an overwhelming quantity of overall material in social media and websites. As the representative example, one of those news items claimed that 5G mobile devices network '*causes Coronavirus by sucking oxygen*

[2] `https://www.enisa.europa.eu/publications/enisa-position-papers-and-opinions/fake-news`
[3] `https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation`

out of your lungs'[4]. Another group said that the virus comes from bat soup[5], while others pointed labs as places where the virus was actually created as part of the conspiracy. According to the *'studies'* published this way, there are also some pieces of advice on how to cure the virus – by gargling vinegar and rosewater, vinegar and salt[6] or – as a classic for plague anxiety – colloidal silver or garlic.

False news may also be the grounds for political agenda, for instance during the 2016 US election. One of the misinformation examples in this campaign was the Eric Tucker's case[7]. Tucker, finding it to be an uncommon occurrence, had photographed a big number of buses that he spotted in the city centre of Austin. Additionally, he watched the announcements concerning the demonstrations in protest at President-elect Donald J. Trump taking place there and arrived at the conclusion that there must have been some connection between both incidents. Tucker tweeted the photos, commenting on them that '*Anti-Trump protestors in Austin today are not as organic as they seem. Here are the busses they came in. #fakeprotests #trump2016 #austin*'. The original post got retweeted at least sixteen thousand times; *Facebook* users shared it over 350,000 times. Later, it turned out that the buses were not involved in the protests in Austin. In fact, they were employed by *Tableau Software*, a company that organised a summit for over 13 thousand people at that time. This resulted in the original post being deleted from Twitter, and instead published the picture of it labelled as '*false*'.

2. Current initiatives worldwide to fight against disinformation

The fake news problem is especially visible in any kind of the social media. In the online report[8] NATO-affiliated researchers claimed that the social media fail to stop the online manipulation. According to the report 'Overall social media companies are experiencing significant challenges in countering the malicious use of their platforms'. First of all, the researchers were easily able to buy tens of thousands of likes, comments and views on Facebook, Twitter, YouTube and Instagram. What is more, Facebook has been recently flooded by numerous fake accounts. It claimed to disable 2.2 billion fake accounts solely in the first quarter of this year! An interesting approach to fake news detecting is to involve the community. That is why in Facebook mobile application reporting tool has become more visible lately.

This is just an example showing how serious and far-reaching the issue may be. The pervasiveness of the problem has already caused a number of fake news prevention initiatives to be developed; they are of both political and non-political character; some are local and some of them are of international scope. The following subsections will present some initiatives of considerable significance.

2.1. Large-scale political initiatives addressing the issue of fake news

From the worldwide perspective, the *International Grand Committee* (*IGC*) *on Disinformation and Fake News* can be considered as the widest present governmental initiative. The *IGC* was founded by the governments of Argentina, Belgium, Brazil, Canada, France, Ireland, Latvia, Singapore, and United Kingdom. Its inaugural meeting[9] was held in the UK in November 2018 and the follow-up ones have been held in Canada (May, 2019) and Ireland (November, 2019). Elected representatives of Finland, Georgia, USA, Estonia, and Australia also attended the last meeting. In addition to general reflections about the topics under analysis, this international board specifically focused on technology and media companies, asking them for liability and accountability. One of the conclusions of the *IGC* latest meeting[10] was that '*global technology firms cannot on their own be responsible in combating harmful content, hate speech and electoral interference*'. As a result, this committee concludes that self-regulation is insufficient.

[4] https://www.mirror.co.uk/tech/coronavirus-hoax-claims-5g-causes-21620766
[5] www.foreignpolicy.com/2020/01/27/dont-blame-bat-soup-for-the-wuhan-virus/
[6] www.bbc.com/news/world-middle-east-51677530
[7] https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html
[8] https://techxplore.com/news/2019-12-nato-social-media.html
[9] https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/declaration-internet-17-19/
[10] https://www.oireachtas.ie/en/press-centre/press-releases/20191107-update-international-grand-committee-on-disinformation-and-fake-news-proposes-moratorium-on-misleading-micro-targeted-political-ads-online/

At the national level, a wide range of actions have been taken. In the case of Australia, interference in political or governmental processes has been one of the main concerns of its Government[11]. As a result, the *Electoral Integrity Assurance Taskforce* (*EIAT*) was created in 2018 to handle risks (cyber interference) to the Australian election system remaining integral. Moreover, in June 2020 the Australian Communications and Media Authority published a paper[12] highlighting that '*48% of Australians rely on online news or social media as their main source of news, but 64% of Australians remain concerned about what is real or fake on the internet*'. The paper discusses potentially harmful effects of fake news or disinformation to users and/or governments at different levels, providing two clear and recent examples which occurred in Australia during the first semester of 2020, such as the bushfire season and the COVID-19 pandemic. In general, two responses to misinformation are pointed out. One that considers international regulatory responses and another one coming from online platforms in terms of how they tackle misconducting users as well as how they address problematic content.

Similarly, in October 2019 the *US Department of Homeland Security* (*DHS*) published a report on *Combatting Targeted Disinformation Campaigns*[13]. In this report, the *DHS* highlights how easy it is nowadays to spread false news through online sources and how '*disinformation campaigns should be viewed as a whole-of-society problem requiring action by government stakeholders, commercial entities, media organizations, and other segments of civil society*'. This report points out the growth on disinformation campaigns since the 2016 US presidential election; at the same time that US and world-wide nations were becoming more aware concerning the potential damage of these campaigns to economy, politics and society in general. Furthermore, better and wider actions are nowadays carried out in real-time, i.e. while the disinformation campaign is ongoing, compared to the first years (until 2018) where most of '*the work on disinformation campaigns was post-mortem, i.e. after the campaign had nearly run its course*'. In this sense, the report summarizes several recommendations to combat disinformation campaigns such as '*government legislation, funding and support of research efforts that bridge the commercial and academic sectors (e.g., development of technical tools), sharing and analysing information between public and private entities, providing media literacy resources to users and enhancing the transparency of content distributors, building societal resilience and encouraging the adoption of healthy skepticism*'. In any case, the importance of '*private and public sector cooperation to address targeted disinformation campaigns on the next five years*' is highlighted.

In Europe, the *EU Commission* has recently updated (July, 2020) a previous report providing a clear set of actions, which are easy to understand and/or apply, in order to fight against fake news and online disinformation[14,15]. The action plan consists of the four following pillars:

1. *Improving the capabilities of Union institutions to detect, analyse and expose disinformation*. This action implies better communication and coordination among the EU Member States and their institutions. In principle, it aims to provide EU members with '*specialised experts in data mining and analysis to gather and process all the relevant data*'. Moreover, it refers to '*contracting media monitoring services as crucial in order to cover a wider range of sources and languages*'. Additionally, it also highlights the need to '*invest in developing analytical tools which may help to mine, organise and aggregate vast amounts of digital data*'.

2. *Stronger cooperation and joint responses to threats*. Since the most significant is the time right after the publishing of the false news, this action aims to have a '*Rapid Alert System to provide alerts on disinformation campaign in real-time*'. For this purpose, each EU Member State should '*designate contact points which would share alerts and ensure coordination without prejudice to existing competences and/or national laws*'.

3. *Enhancing collaboration with online platforms and industry to tackle disinformation*. This action aims to mobilise and provide with an active role to the private sector. Disinformation is most often released in large online platforms owned by the private sector. Therefore, they should be able to '*close down fake accounts active*

[11] https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews

[12] https://www.acma.gov.au/sites/default/files/2020-06/Misinformation\%20and\%20news\%20quality\%20position\%20paper.pdf

[13] https://www.dhs.gov/sites/default/files/publications/ia/ia_combatting-targeted-disinformation-campaigns.pdf

[14] https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation

[15] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56166

on their service, identify automated bots and label them accordingly, and collaborate with the national audio-visual regulators and independent fact-checkers and researchers to detect and flag disinformation campaigns'.

4. *Raising awareness and improve societal resilience*. This action aims to increase public awareness and resilience by activities related to '*media literacy in order to empower EU citizens to better identify and deal with disinformation*'. In this sense, the development of critical thinking and the use of independent fact-checkers are highlighted to play a key role in '*providing new and more efficient tools to the society in order to understand and combat online disinformation*'.

Overall, any activity or action proposed in the above-mentioned cases (*IGC*, Australia, US and EU) could be understood from three different angles: technological, legislative and educative. For instance – technology-wise – the four pillars in the set of actions proposed by the EU Commission are mentioning the use of several kinds of tools (analytical, fact-checkers, etc.), arising as the key component to be considered in current and/or future initiatives. However, from the legislation point of view, pillars 1 and 2 are also showing the need to have a normative framework which facilitates coordination and communication among different countries. Finally – education-wise – pillars 3 and 4 strengthen the importance of media literacy and the development of critical thinking in society. Similarly, activities, actions and recommendations provided by the *IGC*, the Australian government and the *DHS* can be directly linked to these three different concepts (technology, legislation and education). Furthermore, the *BBC* has an available collection of tools, highlights and global media education[16] which could be directly linked to these angles previously described, thus supporting this division into three categories.

The U.S. Department of Defence started to invest about $70 M in order to deploy military technology to detect fake contents as they impact on national security[17]. This *Media Forensics Program* at the *Defence Advanced Research Projects Agency* (*DARPA*) started 4 years ago.

Some efforts are being devoted to the technological component from public bodies. That is the case of the Australian *EIAT*, which was created to provide the *Australian Electoral Commission* with '*technical advice and expertise*' concerning potential digital disturbance of electoral processes[18]. The *Australian Cyber Security Centre* was in charge of providing this technological advice, among other governmental institutions.

In the case of Europe, a high-level group of experts (*HLGE*) was appointed in 2018 by the European Commission to give advice on this topic. Although these experts were not against regulating in some cases, they proposed [14] to mainly take non-regulatory and specific-purpose actions. The focus of the proposal was the collaboration between different stakeholders to support digital media companies in combating disinformation. On the contrary, the *Standing Committee on Access to Information, Privacy and Ethics* (*SCAIPE*) established by the Parliament of Canada suggested in 2019 to impose legal restrictions to media companies in order to be more transparent and force them to remove illegal contents [15]. Similarly, the *Digital, Culture, Media and Sport Committee* (*DCMSC*) created by the Parliament of the United Kingdom strongly encouraged to take legal actions [16]. More precisely, it proposed a mandatory ethical code for media enterprises, to be controlled by an independent body, and forcing these companies to remove those contents to be considered potentially dangerous and coming from proven sources of disinformation. Furthermore, the *DCMSC* suggested to modify laws regarding electoral communications to ensure their transparency in online media.

Regarding electoral processes, it is worth mentioning the Australian initiative; unprecedented offences by foreign interference have been added to the Commonwealth Criminal Code by means of the *National Security Legislation Amendment (Espionage and Foreign Interference) Act 2018*[19]. It defines these foreign offences as the ones that '*influence a political or governmental process of the Commonwealth or a State or Territory or influence the exercise (whether or not in Australia) of an Australian democratic or political right or duty*'.

In the case of India, the Government issued a notice to *Whatsapp* in 2018, because at least 18 people were killed in separate incidents that year after false information was shared through this app[20]. The Minister of Electronics and Information Technology stated that the Indian Government '*was committed to freedom of speech and privacy*

[16] https://www.bbc.co.uk/academy/en/collections/fake-news

[17] https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775

[18] https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id\%3A\%22media\%2Fpressclp\%2F6016585\%22

[19] https://www.legislation.gov.au/Details/C2018C00506

[20] https://www.cbc.ca/news/world/india-child-kidnap-abduction-video-rumours-killings-1.4737041

as enshrined in the constitution of India'. As a result, the posts published to social networks are not subject to governmental regulations. He claimed that '*these social media have also to follow Article 19(2) of the Constitution and ensure that their platforms are not used to commit and provoke terrorism, extremism, violence and crime*'[21]. However, India's Government is working on a new IT Act in order to deploy a stronger framework to deal with cybercrimes[22].

Lastly, it should be mentioned that the third (latest) meeting of the *IGC* was aimed at advancing international collaboration in the regulation of fake news and disinformation. In this sense, experts highlighted that there are conflicting principles regarding the regulation of the internet. This includes the protection of freedom of speech (in accordance with national laws), while simultaneously combating disinformation. Thus, this still is an open challenge.

Finally, from the education perspective, it is worth mentioning that the *EU-HLGE* recommended implementing wide education programs on media and information, in order to educate not only professional users of media platforms but public users in general terms. Similar recommendations were pointed out by the Canadian *SCAIPE*, focusing on awareness-raising campaigns and literacy programs for the whole society.

2.2. Other noteworthy initiatives and solutions

It should also be mentioned that recently some systematic social activity battling with misinformation has appeared and now it is getting more intense. For instance, there is a group of volunteers called Lithuanian '*elves*'. Their main aim is to beat the *Kremlin* propaganda. They scan the social media (*Instagram, Facebook, Twitter*) and report any found fake information on their daily basis.

From the technological perspective, it is worth highlighting that numerous online tools have been developed for misinformation detection. Some available approaches were presented in [17]. This technological development to combat disinformation is led by tech/media companies. This is the case of *Facebook*, that quite recently (May 2020) informed[23] that it is combating fake news by means of its *Multimodal Content Analysis Tools*, in addition to 35 thousand human supervisors. This AI-driven set of tools is been applied to identify fake or abusive contents related to Coronavirus. The image-processing system extracts objects that are known to violate its policy. Then the objects are stored and searched in new ads published by users. *Facebook* claims that this solution, based on supervised classifiers, does not suffer from the limitations of similar tools when facing images created by common adversarial modification techniques.

In the *BlackHat Europe 2018* event that was held in London, the *Symantec Corporation* displayed its demo of a deepfake detector[24]. In 2019, Facebook put 10 million dollars[25] into the *Deepfake Detection Challenge* [18] aimed at measuring progress on the available technology to detect deepfakes. The best model (in terms of precision metric for published data) that won this contest, performed quite poorly (65.18% precision) when validated with new data. This means that it still is an open challenge and a great research effort is still required to get robust fake detection technology. In addition to these huge companies, some other startups are developing anti-fake technologies, such as the *DeepTrace* based in Netherlands. This company aims at building the '*antivirus for deepfakes*'[26].

Some other technological projects are being run at present time; the *AI Foundation* raised 10 million dollars to develop the *Guardian AI* technologies, a set of tools comprising *Reality Defender*[27]. This intelligent software is intended to support users in identifying fake contents while consuming digital resources (such as web browsing). No further technical details are available yet.

[21] https://www.thehindu.com/news/national/fake-news-safety-measures-by-whatsapp-inadequate-says-ravi-shankar-prasad/article24521167.ece

[22] https://www.thehindu.com/business/Industry/centre-to-revamp-it-act/article30925140.ece

[23] https://spectrum.ieee.org/view-from-the-valley/artificial-intelligence/machine-learning/how-facebook-is-using-ai-to-fight-covid19-misinformation

[24] https://i.blackhat.com/eu-18/Thu-Dec-6/eu-18-Thaware-Agnihotri-AI-Gone-Rogue-Exterminating-Deep-Fakes-Before-They-Cause-Menace.pdf

[25] https://www.reuters.com/article/us-facebook-microsoft-deepfakes/facebook-microsoft-launch-contest-to-detect-deepfake-videos-idUSKCN1VQ2T5

[26] https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/will-deepfakes-detection-be-ready-for-2020

[27] https://aifoundation.com/responsibility/

3. A systematic overview of ML approaches for fake news detection

A comprehensive, critical analysis of previous ML-based approaches to false news detection is presented in the following part of the paper. As already mentioned in Section 1, and as graphically presented in Fig. 1, these methods can analyze different types of digital content. According to that, in this section we will overview the methods for (*i*) text-based and Natural Language Processing (NLP) analysis, (*ii*) reputation analysis, (*iii*) network analysis, and (*iv*) image-manipulation recognition.

3.1. Text analysis

Intuitively, the most obvious approach to automatically recognizing fake news is NLP. Despite the fact that the social context of the message conveyed in electronic media is a very important factor, the basic source of information necessary to build a reliable pattern recognition system is the extraction of features directly from the content of the analyzed article. Several main trends may be distinguished within the works carried out in this area. Theoretically, the simplest is the analysis of text representation without linguistic context, most often in the form of *bag-of-words* (first mentioned in 1954 by Harris [19]) or *N-grams*, but also the analysis of psycholinguistic factors, syntactic and semantic analysis are commonly used.

3.1.1. NLP-based data representation

As it is rightly pointed out by Saquete et al. [20], each of the subtasks distinguished within the domain of detecting false news is based on the tools offered by NLP. Basic solutions are built on *bag-of-words*, which counts the occurrences of particular terms within the text. A slightly more sophisticated development of this idea are *N-grams*, which tokenize not individual words, but their sequences. The range of the definition allows to define *bag-of-words* as *N-grams* with *n* equal to one, making them *unigrams*. It has also to be noted that the sheer number of *N-grams* in each document is strongly dependent on its length and for the needs of the construction of pattern recognition systems it should be normalized to the document (*Term frequency* : TF [21]) or to the set of documents used in the learning process (*Term frequency - inverse document frequency : TF-IDF* [22]).

Despite the simplicity and age of these solutions, they are successfully utilized in solving the problem of detecting false news. A good example of such application is the work of Hassan et al. [23], comparing the effectiveness of *Twitter* disinformation classification using five base classifiers (*Lin-SVM, RF, LR, NB and KNN*), comparing them with methods of TF and TF-IDF attribute extraction with *N-grams* of various lengths. On the basis of the PHEME [24] dataset, they showed the effectiveness of methods using simultaneously different lengths of word sequences, combining *unigrams* with *bigrams* in the extraction. A similar approach is a common starting point for most analyzes [25, 26], enabling further suggestions for considerations, through the in-depth review of base classifiers [27] or the use of ensemble methods [28].

Other interesting trends within this type of analysis include the detection of unusual tokens, for example text elements, insistently repeated denials and curses, or question marks, emoticons and multiple exclamation marks, that are most often rejected at the stage of preprocessing [29, 30, 31]. As in many other application fields, deep neural networks are a promising branch of Artificial Intelligence. Hence, they are also playing a role here, being a popular alternative to classic models [32].

3.1.2. Psycholinguistic features

The psycholinguistic analysis of texts published on the Internet is particularly difficult due to the limitation of messages to their verbal part only and the peculiar characteristics of such documents. Existing and widely cited studies [33, 34] allow us to conclude that the messages that try to mislead us are characterized, for example, by an enormous length of some sentences they contain along with their lexical limitation, increased repetition of key theses or a reduced formality of the language used. These are often extractable and measurable factors that can be more or less successfully used in the design of the pattern recognition system.

An interesting work giving a proper overview on psycholinguistic data extraction is the detection of character assassination attempts by troll actions, performed by El Marouf et al. [35]. Six different feature extraction tools were used in the construction of the dataset for supervised learning. The first is *Linguistic Inquiry and Word Count* (*LIWC*) [36], which in its latest version allows to obtain 93 individual characteristics of the analyzed text, including both simple measurements, like the typical word count within a given phrase and complex analysis of the grammar

used or even the description of the author's emotions or the cognitive processes performed by them. It is a tool that has been widely used for many years in a variety of problems ([37], [38], [39], [40]), fitting well for detecting fake news. Notwithstanding, in [41] authors presented a sentiment analysis proposal that classifies the sample text as irony or not.

Another tool are *POS* Tags, assigning individual words to *Parts-of-speech* and returning their percentage share in the document [42]. The basic information about the grammar of the text and the presumed emotions of the author obtained in this way are supplemented with the knowledge acquired from *SlangNet* [43], *Colloquial WordNet* [44], *SentiWordNet* [45] and *SentiStrength* [46], returning, in turn, information about slang and colloquial expressions, indicating the author's sentiment and defining it as positive or negative. The system proposed by the authors, using 19 of the available attributes and *Multinomial Naive Bayes* as a base classifier, allowed to obtain a 90% score in the *F-score* metric.

An extremely interesting trend in this type of feature extraction is the use of behavioral information that is not directly contained in the entered text, but can be obtained by analyzing the way it was entered [47, 48].

3.1.3. Syntax-based

The aforementioned methods of obtaining useful information from the text were based on the analysis of the word sequences present in it or the recognition of the author's emotions hidden in the words. However, a full analysis of *natural language* requires an additional aspect in the form of processing the syntactic structure of expressed sentences. Ideas expressed in simple data representation methods, such as *N-grams*, were developed to *Probability Context Free Grammars* [49], building distributed trees describing the syntactic structure of the text.

In syntactic analysis, we do not have to limit ourselves to the structure of the sentence itself, and in the case of social networks, extraction can take place by building the structure of the whole discussion, as Kumar and Carley show [50]. Extraction of this type [51] seems to be one of the most promising tools in the fight against fake news propagation [52].

3.1.4. Non-linguistic methods

The fake news classification is not limited to the linguistic analysis of documents. The studies analysing different kinds of attributes which may be of use in the same setting [53] are interesting. Amidst the typical methods that the study comprises, there are the analyses of the creator and reader of the message, as well as the contents of the document and its positioning within social media outlets being verified [54]. Another method which shows promise is analysing images; this approach concerns fake news in the form of video material [55]. Similarly, the study by [3] is evenly thought-provoking; it proposes to divide the methods of linguistic and social analyses. The former group of models encompasses the semantic, rhetorical, discourse and simple probabilistic recognition ones. The latter set comprises analysing how the person conveying the message behaves in social media and what context their entries are building. Then, [30] has based the design of the recognition models on the behavior of the authors, where the background of the post (both the posted and forwarded ones) does depend on their bodies, and at the same time refers to other texts. Diverse representations of data were analysed by [56], whilst [57] has examined various variants of stylometric metrics.

Numerous issues relating to fake news detection have been studied by [5] and indicated that it is possible to apply the *Scientific Content Analysis* (SCAN) approach in order to tackle the matter. In [58], an effective method was advanced which aims at verifying the news automatically. This approach would depart from analysing texts and head for image data. On the other hand, [54] suggests performing analyses of the posts from social networks within the context of data streaming, arguing that this approach addresses their dynamic character. Support Vector Machine (SVM) has been utilised by [29] in order to recognize which messages are genuine, false or of satirical nature. A similar type of classifier has been employed by [59] as well; in their work, semantic analysis as well as behavioural feature descriptors are to uncover the media entries which may be false.

The comparison was made by [60] so as to assess a number of classification methods which base on linguistic features. According to its outcomes, successful detection of false information can base on the already familiar classifier models (particularly ensembles). In [61], it has been indicated that the issue of detecting false information is generally limited to the classification task, although anomaly detection and clustering approaches might be utilised for it. Lastly, in [62], the NLP tools-based method was proposed to be applied for analysing Twitter posts. According

to this approach, each post was assigned credibility values owing to the fact that the researchers viewed this issue as a regression task.

3.2. Reputation analysis

The reputation[28] of an individual, a social group, a company or even a location, is understood as the estimation of it; usually it results from the determined criteria which influence social assessment. Within a natural society, reputation is the mechanism of social control, characterised by high efficiency, ubiquity and spontaneity. Social, management and technological sciences aim at investigating this matter. It must be acknowledged that reputation influences communities, markets, companies, and institutions alike; in other words, it has an influence over both the competitive and cooperative contexts. Its influence may extend as far as the relations between the whole countries; the idea's importance is appreciated in politics, business, education, online networks, and diverse, innumerable domains. Thus, reputation can be regarded to be reflecting the identity of a given social entity.

In technological sciences, the reputation of a product or a site often needs to be measured. Therefore a reputation score, which represents it in a numeric manner, is computed. The calculation may be performed by means of a centralized reputation server or distributively, by means of local or global trust metrics [63]. This evaluation may be of use when supporting entities in their decisions concerning if they should rely on someone or buy a given product, or not.

The general concept behind reputation systems is allowing the entities evaluate one another or assess an object of their interest (like articles, publishers, etc.); then subsequently utilize the collected evaluations to obtain trust or reputation scores, concerning the sources and items within the system. In order for systems to act in this way, reputation analysis techniques are used, which actually support the ability to establish in an automatic manner the way various keywords, phrases, themes or contents created by users are analysed amongst the mentions of diverse origin. More specifically, in the news industry, there are two kinds of sources of reputation for an article/publisher [64]; reputation from content and reviews/feedback on the one hand, and reputation from IP or domain on the other one.

Given the technological advances, all types of data can be collected: type of comments, their scope, keywords, etc. Especially in the news industry, there are a few key characteristics that can differentiate a trustworthy article from a fake one. A common characteristic is the anonymity fake news publishers choose behind their domain. The results from the survey in [64] showed that whenever the person publishing contents wishes to protect their anonymity, the online who-is information will indicate the proxy as the organization that registered it. On the other hand, the renowned, widespread newspapers usually prefer to register under their actual company names. Another indicative feature for fake news is the duration of time that publishers spend on the websites with the intention of disseminating false information. Often, it is rather brief in comparison to the real news publishers. Domain popularity is also a good indicator regarding the reputation, as it measures the views of the website gets every day, per a visiting person. It seems logical that a well-know website features a greater number of views per person, as they have the tendency to devote more time to surfing the contents and looking at the various sub pages. It has been indicated in [64] that the domains of the trustworthy web pages which post genuine information are far more popular than the ones which disseminate untruth. The reason for this is that normally the majority of web pages that publish false news either stop publishing news very soon, or the readers spend much less time browsing those sites.

So, reputation scoring ranks fake news based on suspicious domain names, IP addresses and review/feedback by the readers by providing a measure that indicates whether the specific website is high or low on reputation. Different techniques for reputation scoring have been proposed in the literature. [65] describes the application of the *Maximum Entropy Discrimination* (MED) classifier. It is utilized to score the reputation of web pages. It is done on the basis of the information which creates a reputation vector. It includes multiple factors for the online resource, such as the state in which the domain was registered, the place where the service is hosted, the place of an internet protocol address block and when the domain was registered. It also considers the popularity level, IP address, how many hosts there are, top-tier domain, a number of run-time behaviours, *JavaScript* block count, the number of images, immediate redirect, and response latency. The paper by [66] also relates to the issue. The scientists have created the *Notos* reputation system which makes use of the unique DNS characteristics in filtering out the malevolent domains on the basis of their past engagement in harmful or genuine online services. For every domain, the authors utilised

[28] `https://www.definitions.net/definition/REPUTATION`

clustering analysis of network-based, zone-based and evidence-based features, so as to obtain the reputation score. As the majority of the methods do not perform the scoring online, so it may use processing-intensive approaches. The assessment that used real-world data, including the traffic from vast ISP networks, has proved that *Notos*'s accuracy in recognizing emerging, malevolent domains in the DNS query traffic that was monitored was indeed very high, with the true positive score of 96.8% and false positive one - 0.38%.

According to the above mentioned, a reputation analysis system is actually a cross-checking system that examines a set of trustworthy databases of blacklists in an automatic manner, usually employing ML techniques that recognize malicious IP addresses and domains based on their reputation scores. More specifically, in [67], a ML model relying on a deep neural architecture which was trained using a big passive DNS database is presented. *Mnemonic*[29] supplied the passive DNS data utilised for the sake of the paper. The raw dataset consisted of 567 million aggregated DNS queries, gathered in the span of almost half a decade. The variables which define every entry are as follows: the type of a record, a recorded query, the response to it, a *Time-to-Live* (TTL) value for the query-answer pair and a timestamp for when the pair occurred at first, as well as the total number of instances when the pair occurred, within a given period. The method is capable of pinpointing 95% of the suspicious hosts, the false positive rate being 1:1000. Nevertheless, the amount of time required to train turned out to be exceptionally high because of the vast amount of the data needed for it; the delay information has not been assessed.

Segugio, an innovative defense system based on behaviour, is introduced in [68]. It makes it possible to track the appearance of newly-appearing, malware-control domain names within huge ISP networks in an efficient manner, with the true positive rate (TP) reaching 85%, and false positive rate (FP) lower than 0.1%. Nevertheless, both TP and FP have been counted on the basis of 53 new domains; proving the correctness based on such a little set may be a challenging task.

Lastly, in [69], an innovative novel granular SVM is presented, namely a boundary alignment algorithm (GSVM-BA), which repeatedly eliminates the positive support vectors from the dataset used for training, in order to find the optimum decision boundary. To accomplish it, two groups of feature vectors are extracted from the data; they are called the breadth and spectral vectors.

3.3. Network data analysis

Network analysis refers to the Network theory, which studies graphs, being a representation of either symmetric or asymmetric relations between discrete objects. This theory has been of use in various fields including statistical physics, particle physics, computer science, electrical engineering, biology, economics, and others. The possible applications of the theory comprise the World Wide Web (WWW), Internet, as well as logistical, social, epistemological and gene regulatory networks, etc. In computer/network sciences, the network theory belongs to graph theory. This means that one may define a network as a graph, where nodes and edges have their attributes (like names).

Network-based detection of false news applies the data on the social context uncovered in news propagation. Generally speaking, it examines two kinds of networks, namely the homogeneous and heterogeneous ones. Homogeneous networks (such as Friendship, Diffusion, and Credibility networks, etc.) contain one kind of nodes and edges. For instance, in credibility networks, people present their points of view regarding the original news items by means of social media entries. In them, they might either share the same opinions (which thus support one another), or conflicting opinions (this is turn may lower their credibility scores). If one were to model the aforementioned relations, the credibility network may be applied to assess the level of veracity of the news pieces, by means of the credibility scores of every particular social network entry (related to the news item) being leveraged. On the other hand, heterogeneous networks are characterised by having numerous kinds of nodes or edges. The main benefits they bring is the capability of representing and encoding the data and relations from various positions. Some well-known networks used for detecting false information are Knowledge, Stance, and Interaction Networks. The first type incorporates linked open data, like DBdata and Google Relation Extraction Corpus (GREC), as a heterogeneous network topology. When inspecting for fact by means of a knowledge graph, it is possible to verify if the contents of the news items may be gathered from the facts that are present in the knowledge networks, whilst Stances (viewpoints) represent people's attitudes to the information, i.e. in favour, conflicting, etc, [70].

[29] `https://www.mnemonic.no/`

In detecting false news, network analysis is performed in order to evaluate the truth value of the news item; one may formalize it as a classification issue which needs obtaining relevant features and building models. As part of feature extraction, the differential qualities of information items become captured in order to create efficient representations; on the basis of them, several models are created, for learning at transforming the features. Contemporary advancements in network representation learning, e.g. network embedding and deep neural networks, let one apprehend the features of news items in an enhanced manner, from supplementary data like friendship networks, temporal user engagements, and interaction networks. Moreover, knowledge networks as secondary data may make it easier to challenge the truthfulness of the news pieces by means of network pairing operations, including path finding and optimizing the flow.

Across network levels, the data from the social media concerning propagating the news and its spreaders has not been examined to a significant degree, yet. In addition to this, it has not been much used in an explainable manner for detecting false information, too. The authors of [71] have suggested a network-based pattern-driven model for detecting false information; it proved robust against the news items being manipulated by malicious actors. The model makes use of patterns in disseminating fake data within social networks, as it turns out that, in comparison with the true ones, false news is able to (*i*) disseminate further and (*ii*) engage a larger number of spreaders, where they oftentimes prove to be (*iii*) more fully absorbed in the news and (*iv*) more closely linked within the network. The characteristics which represent the aforementioned patterns have been developed at several network levels (that is, node-, ego-, triad-, community-, and network-level), that may be utilised in a ML supervised-learning framework for false news detection. The patterns involved in the mentioned study regarding fake news concern the spreading of the news items, the ones responsible for doing it, and the relations among those spreaders. Another example of network analysis in the false news detection may be found in [72], where a framework is proposed which uses a tri-relationship model (TriFN) amongst the news article, spreader and publisher. For such a network, a hybrid framework is composed of three major parts which contribute to detecting false news: (*i*) entity embedding and representation, (*ii*) relation modeling, and (*iii*) semi-supervised learning. The actual model possesses four meaningful parameters. α and β manage the inputs from social relationship and user-news relations. γ manages the input of publisher partisan and η controls the input provided by semi-supervised classifier. According to [72], TriFN is able to perform well whilst detecting, at the initial stages of disseminating news.

3.4. Image based analysis and detection of image manipulations

In the last decade, digital images have thoroughly replaced conventional photographs. Currently it is possible to take a photo using not only cameras but also smartphones, tablets, smart watches and even eye glasses. Thus, thousands of billions of digital photos are taken annually. The immense popularity of image information fosters the development of the tools for editing it, for instance *Photoshop* or *Affinity Photo*. The software lets users manipulate real-life pictures, from low-level (adjusting the brightness in a photo) to high-level semantic contents (replacing an element of a family photograph). Nonetheless, the possibilities provided by the photo manipulation tools may be seen as a double-edged sword. It enables making the pictures more pleasing to the eye, and also inspires users in their expressing and sharing their visions on visual arts; however, the contents of the photo may be forged more easily, with no visible hints left. Thus, it makes it easier to spread fake news. Hence, with the passage of time, several scientists have developed the methods for detecting photo tampering; the techniques concentrate on copy-move forgeries, splice forgeries, inpainting, image-wise adjustments (such as resizing, histogram equalization, cropping, etc.) and other ones.

So far, numerous researchers have presented some approaches for detecting fake news and image tampering. In [73], the authors proposed an algorithm which is able to recognise if the picture has been altered and where, taking advantage of the characteristic footprints that various camera models leave in the images. Its way of working is based on the fact that all the pixels of the pictures that have not been tampered with ought to appear as if they had been taken by means of a single device. Otherwise, if the image has been composed of multiple pictures, then the footprints left by several devices may be found. In the presented algorithm, a *Convolutional Neural Network* (CNN) is exploited for obtaining the characteristics which are typical of the specific camera model, from the studied image. The same algorithmic was also used in [74].

Authors in [75] used *Structural Similarity Index Measure* (SSIM) instead of more classical approaches to modification detection techniques such as: *mean squared error* (MSE) and *peak signal to noise ratio* (PSNR). Moreover,

they moved the whole solution to cloud computing that is claimed to provide security, the information being deployed quicker, as well as the data being more accessible and usable.

The technique presented in [76] uses the chrominance of the *YCbCr* colour space; it is considered to be able to detect the distortion resulting from forgery more efficielty than all the further colour components. When extracting features, selected channel gets segmented into blocks which overlap. Thus, *Otsu-based Enhanced Local Ternary Pattern* (OELTP) has been introduced; it is an innovative visual descriptor aimed at extracting features from the blocks. It extends the *Enhanced Local Ternary Pattern* (ELTP) which recognises the neighbourhood on a range of three values (-1, 0, 1). Subsequently, the energy of OELTP features gets assessed in order to decrease the dimensionality of features. Then, the sorted characteristics are used for training the SVM classifier. Lastly, the image is labelled, either as genuine or manipulated.

The colour space *YCbCr* was also used in [77]. The presented approach takes advantage of the fact that during when being merged, at least two pictures get engaged in copying and pasting. In case of tampering with JPEG images, the forgery might not follow the same pattern; a piece of an uncompressed picture may be pasted into a compressed JPEG file or the other way round. Such manipulated images being re-saved in JPEG format with different quality factors can possibly result in the emergence of double quantization artefacts in DTC coefficients.

According to the method proposed in [78], the input file becomes pre-processed by converting its colour space from RGB to grey level; following that, the image features (*Histogram of Oriented Gradient* (HOG), *Discrete Wavelet Transform* (DWT) and *Local Binary Patterns* (LBP)) get distilled from the grey level colour space. This fitting group of features is merged to create a feature vector. The *Logistic Regression* classifier is utilised to construct a model and to discriminate a manipulated, authenticated picture. The suggested technique enhances the correctness of detection by applying combined spacial features, that is the spatial- and frequency-based ones.

Two types of features were also used in [79]. In this work, the authors decided to convert the source image into grey scale and use *Haar Wavelet Transform*. Then, the vector features are calculated using HOG and *Local Binary Patterns Variance*. The classification step is founded upon the Euclidean distance.

Within [80], *Speeded Up Robust Features* (SURF) and *Binary Robust Invariant Scalable Keypoints* (BRISK) descriptors were used in order to reveal and pinpoint single and multiple copy–move forgeries. The features of SURF prove robust against diverse post-processing attacks, like rotation, blurring and additive noise. Nevertheless, it seems that features of the BRISK are equally as robust in relation to detecting the scale-invariant forged regions, along with the poorly localized keypoints of the objects within the forged image.

However, the *state-of-the-art* techniques deal not only with copy-move forgeries, but with other types of modifications, too. The paper [81] presents the contrast enhancement detection based on numerical measures of images. The proposed method includes division in non-overlapping blocks and then, mean, variance, skewness and kurtosis of block calculation. The method also uses DWT coefficients and SVM for original/tampered classification. Whereas, in [82] authors claimed that real and fake colorized images can differ in *Hue* and *Saturation* channels of *Hue-Saturation-Value* (HSV) colour space. Thus, the presented approach using histogram equalisation and some other statistical features enables authenticity verification in the colorizing domain.

Overview of all the tools of feature extraction from fake news mentioned in this section is presented in Table 1.

4. Research interest and popular datasets used for detecting fake news

4.1. Is fake news an important issue for the research society?

The raising interest in the fake news detection domain might easily be noticed by checking how many scientific articles have concerned this topic, according to commonly-used and relevant databases. Such metrics are shown in Figure 4 that depicts the number of publications on fake news detection per year and database. According to that, in the *Scopus* database there are 5 articles associated to the '*fake news detection*' keyword and published in 2016, 44 in 2017 , 150 in 2018 and 371 in 2019. In the *Web of Science* database there are 4 articles published in 2016, 24 in 2017, 62 in 2018 and finally 86 in 2019. There is a similar when looking at the *IEEExplore* database, stating that there were 3, 16, 59 and 133 articles published respectively in 2016, 2017, 2018 and 2019.

In addition to the published papers, another key metric to check the interest of the research community and funding agencies on a certain topic is the number of funded projects in competitive calls. According to this idea, a list of the

Table 1: Overview of ML extractors for fake news detection.

CATEGORY	EXTRACTOR	CONTEXT OF EXTRACTION
NLP	*N-grams*	Primitive tool of tokenization.
	TF	Normalization of tokens to the documents.
	TF-IDF	Normalization of tokens to the corpora.
PSYCHOLINGUISTIC	*LIWC*	Collection of 93 individual characteristics from word counts to grammar analysis.
	POS Tags	Assignment of words to parts of speech.
	SlangNet	Various models of recognition of slang and colloquial expressions.
	ColloquialWordNet	
	SentiWordNet	
	SentiStrength	
SYNTAX	*PCFG*	Construction of distributed trees describing dynamic structure of text.
	Tree LSTMs	Application of Long Short-Term Memory networks to analysis of content distribution.
NON-LINGUISTIC	*SCAN*	Scientific content analysis.
REPUTATION	*MED*	Reputation vector of a publisher.
	Notos	Reputation vector from DNS characteristic.
	Segugio	Tracking the appearance of new, malware-control domain names within huge ISP networks.
IMAGE	*CNNs*	Primitive tool of feature extraction of digital signals.
	SSIM	Modification detection metric.
	OELTP	Visual block descriptor.
	HOG, DWT, LBT	Primitive tools of feature extraction.
	SURF, BRISK	Detection of copy-move forgeries.

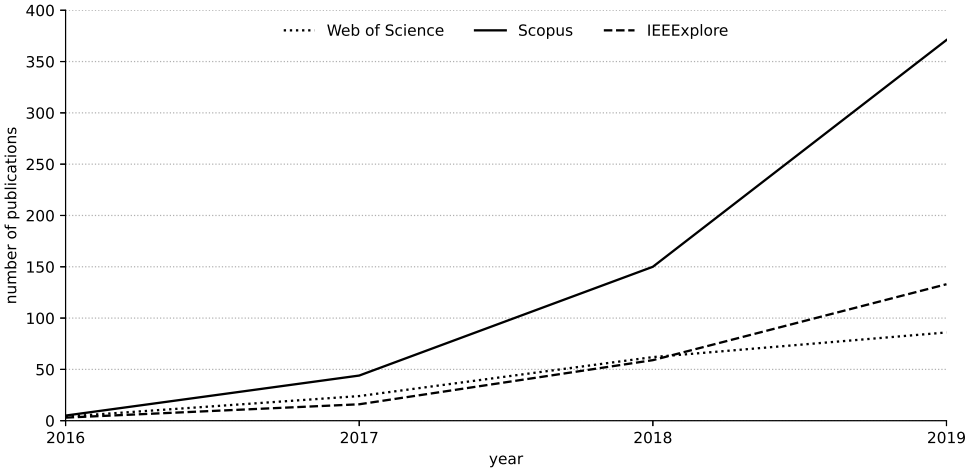

Figure 4: Evolution of the number of publications *per year* retrieved from the keyword "*fake news detection*" according to *Web of Science*, *Scopus* and *IEEExplore*.

research EU-funded projects can be seen in Table 2. Data have been compiled from CORDIS[30] database (July 2020)

[30] `https://cordis.europa.eu/projects/enl`

searching the terms: '*fake news*', '*disinformation*', and '*deepfake*'.

Table 2: EU-funded research projects. Data extracted from CORDIS database in July 2020.

YEAR	NUMBER	PROJECT ACRONYMS
2014	1	PHEME
2015	0	—
2016	5	COMPROP, DEBUNKER, DANTE, ENCASE, *InVID*
2017	1	BOTFIND
2018	7	DYNNET, FANDANGO, *GoodNews*, JOLT, *SocialTruth*, SOMA, *WeVerify*
2019	8	*Factmata*, FAKEOLOGY, DIGIACT, *Media and Conspiracy*, NEWTRAL, QI, RUSINFORM, TRUTHCHECK
2020	5	DICED, *mistrust*, *News in groups*, PRINTOUT, RADICALISATION

Among other EU projects, the *Social Truth* project can be highlighted. It is a project funded by the *Horizon 2020* RD program and it addresses the burning issue of fake news. Its purpose is to deal with this matter in a way that would let vendors to lock-in the solution, build up the trust and reputation by means of the *blockchain* technology, incorporate *Lifelong Learning Machines* which are capable of spotting the paradigm changes of false information and provide a handy digital companion which would be able to support the individuals in their verifying the services they use, from within web browsers they utilise. To meet this objective, ICT engineers, data scientists and *blockchain* experts belonging to both the industry and the academia, together with end-users and use-cases providers have created a consortium in order to combine their efforts. Further details may be found in [83].

It can be concluded that the European Union (EU) works hard in order to combat online misinformation and to educate the society. As a result, increasing amounts of economic resources are being invested.

4.2. Image tampering datasets

There are multiple datasets of modified images available online. One of them is CASIA dataset (in fact, two version of this dataset: CASIA ITDE *v1.0* and CASIA ITDE *v1.0*) described in [84] and [85]. The ground truth input pictures are sourced from the CASIA ITDE image tampering detection evaluation (ITDE) *v1.0* database; it comprises of images belonging to eight categories (animal, architecture, article, character, nature, plant, scene and texture), sized 384x256 or 256x384. Comparing datasets CASIA ITDE *v1.0* and CASIA ITDE *v2.0*, the newer one proves to be more challenging and comprehensive. It utilises post-processing, such as blurring or filtering of the tampered parts to render the manipulated images seem realistic to one's eye. In the CASIA ITDE *v2.0* dataset, there can be a number of tampered versions for each genuine image.

In accordance with CASIA ITDE *v2.0*, the manipulated images are created by applying crop-and-paste operation in *Adobe Photoshop* on the genuine pictures, and the altered areas might be of irregular shapes and various measurements, rotations or distortions.

Among datasets of modified images available online, the one proposed in [86] should also be enumerated. It contains unmodified/original images, unmodified/original images with JPEG compression, 1-to-1 splices (i.e. direct copy of snippet into image), splices with added Gaussian noise, splices with added compression artefacts, rotated copies, scaled copies, combined effects and copies that were pasted multiple times. Mostly, the subsets exist in downscaled versions as well. There are two image formats available: JPEG and TIFF, even though the TIFF format can have a size up to 30 GB.

The other dataset was provided by *CVIP Group* working at *Department of Industrial and Digital Innovation* (*DIID*) of University of Palermo [87]. The Dataset comprises medium-sized images (most of them 1000x700 or 700x1000) and is further divided into multiple datasets (*D0, D1, D2*). The first dataset *D0* is composed of 50 not compressed images with simply translated copies. For remaining two sets of images (*D1, D2*), 20 not compressed images were selected, showing simple scenes (single object, simple background). *D1* subset has been made by copy-pasting rotated elements, while *D2* - scaled ones.

Next dataset was proposed in [88]; it is the *CG-1050* dataset which comprises 100 original images, 1050 tampered images and their corresponding masks. The dataset is divided into four directories: original, tampered and mask images, along with a description file. The directory of original images contains 15 colour and 85 grayscale images.

The directory of tampered images comprises 1050 images obtained by one of the following methods of tampering: copy-move, cut-paste, retouching and colorizing.

There are also some datasets consisting on videos. As an example, the *Deepfake Detection Challenge Dataset* can be listed [18]. This dataset contains 124k videos that have been modified using eight facial modification algorithms. The dataset is useful for deepfake modification of videos.

4.3. Fake news datasets

The LIAR dataset, where there are almost thirteen thousand manually labelled brief statements in varied context taken from the website *polifact.com*, was introduced in [89]. It contains the data collected over a span of a decade and marked as: *pants-on-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*. The label distribution is fairly well-balanced: besides 1,050 *pants-on-fire* cases, there are between 2,063 to 2,638 examples for each label.

The dataset used in [29] in fact consisted of three datasets: (*i*) *Buzzfeed* – data collected from *Facebook* concerning the US Presidential Election, (*ii*) *Political news* – news taken from trusted sources (*The Guardian*, *BBC*, etc.), fake sources (*Ending the Fed*, *Inforwars*, etc.) and satire (*The Onion*, *SatireWire*, etc.) and (*iii*) *Burfoot and Baldwin* – the dataset proposed in 2009, containing mostly real news stories. Unfortunately, the whole dataset is not well-balanced. It contains 4,111 real news, 110 fake news and 308 satire news.

The CREDBANK dataset was introduced in [90]. It is an extensive, crowd-sourced dataset containing about 60 million tweets covering 96 days, beginning from October 2015. The tweets relate to more than 1,000 news events, with each of them checked for credibility by 30 editors from *Amazon Mechanical Turk*.

The *FakeNewsNet* repository, which is updated in a periodical manner, was proposed by the authors of [91]. This dataset contains the combination of news items (source, body, multimedia) and social background details (user profile, followers/followee) concerning fake and truthful materials, gathered from *Snopes* and *BuzzFeed*, that have been reposted and shared on *Twitter*.

The next dataset is *ISOT Fake News Dataset* that was described in [48]. It is very well-balanced: contains over 12,600 false and true news items each. The dataset was gathered using real world outlets; the truthful items were collected by crawling the articles from *Reuters.com*. Conversely, the articles containing false information were picked up from diverse sources. The false news articles were gathered from unreliable web pages flagged by *Politifact* (a US-based fact-checking entity) and *Wikipedia*.

Another method for detecting fake news, called the multimodal one, was shown in [92]. There, authors defined possible features that may be used during the analysis. They are: textual features (statistical or semantic), visual features (image analysis) and social context features (followers, hashtags, retweets). In the presented approach, textual and visual features were used for detecting fake news.

5. Conclusions, further challenges and way forward

This section draws the main conclusions of the present research, regarding the application of advanced ML techniques. Additionally, open challenges in the disinformation arena are pointed out.

5.1. Streaming nature of fake news

It should be underlined that most of the papers addressing fake news detection ignore the streaming nature of this task. The profile of items labelled as fake news might shift over time because the spreaders of false news are conscious of the fact that automatic detection systems could detect them. As a result, they try to avoid their messages being identified as fake news by changing some of their characteristics. Therefore, in order to continuously detect them, ML-driven systems have to react to these changes, known as *concept drift* [93]. It requires to equip the detection systems with mechanisms able to adapt to changes. Only a few papers have attempted to develop fake news detection algorithms taking into consideration the streaming nature of the data [28]. Though a number of researchers noted social media should be considered as data streams, only Wang and Terano [94] used appropriate techniques for data stream analysis. Nevertheless, their method is restricted to quite short streams and probably did not reflect the non-stationary nature of the data. Ksieniewicz et al. [95] employed NLP techniques and treated incoming messages as a non-stationary data stream. The computer experiments on real-life false news datasets prove the usefulness of the suggested approach.

Table 3: The review of existing datasets.

DATASET	ELEMENTS		CITATION
	QUANTITY	CATEGORY	
LIAR	1,050	*pants-on-fire*	[89]
	2,063 – 2,638	*others*	
Buzzfeed dataset	4,111	*real news*	
+ Political news dataset	110	*fake*	[29]
+ Burfoot and Baldwin dataset	308	*satire*	
CREDBANK	*60 million tweets*		[90]
FakeNewsNet	*no data*		[91]
ISOT Fake News Dataset	> 12,600	*real news*	[48]
	> 12,600	*fake*	
Twitter + Weibo	12,647	*real news*	[92]
	10,805	*fake*	
	10,042	*images*	

5.2. Lifelong learning solutions

Lifelong ML systems may transcend the restrictions of canonical learning algorithms that require a substantial set of training samples and are fit for isolated single-task learning [96]. Key features which should be developed in the systems of this kind in order to take advantage of prior learned knowledge comprise feature modeling, saving what had been learnt from past tasks, transferring the knowledge to upcoming learning tasks, updating the previously learnt things and user feedback. Additionally, the idea of a '*task*' which is present in several conventional definitions [97] of lifelong ML models, proves difficult to specify in numerous real-life setups (oftentimes, it seems hard to tell when a given task ends and the subsequent one begins). One of the major troubles is the dilemma of *stability and plasticity*, i.e. the situation where the learning systems must compromise between learning new information without forgetting the previous one [98]. It is visible in the catastrophic forgetting phenomenon, which is described as a neural network forgetting the previously learned information entirely, after having been exposed to new information.

We believe that lifelong learning systems and methods would perfectly fit the fake news problem where content, style, language and types of fake news change rapidly.

5.3. Explainability of ML-based fake news detection systems

Additional point which must be considered at present time is the explainability of ML and ML-based fake news detection methods and systems. Unfortunately, numerous scientists and systems architects utilise deep-learning capacities (along with other black-box ML techniques) in performing detecting or prediction assignments. However, the outcome produced by the algorithms is given with no explanation. Explainability concerns the extent to which a human is able to comprehend and explain (in a literal way) the internal mechanics driving the AI/ML systems.

Indeed, for the ML-based fake detection methods to be successful and widely trusted by different communities (journalism, security etc.), the relevant decision-makers in a realistic environment need the answer to the following question: what is the reason for the system to give certain answers [99]?

5.4. The emergence of deepfakes

It is worth mentioning that, going one step further in this subject, a new phenomenon has recently appeared, referred to as *deepfakes*. Initially, they could be defined as hyper-realistic movie files applying face swaps which do not leave much trace of having been tampered with [100]. This ultimate manipulation now consists in the generation of fake media resources by using AI face-swap technology. The contents of graphical deepfakes (both pictures and videos) are mainly the people whose faces are substituted. On the other hand, there are also deepfakes recordings in

which the voices of people are simulated. Although there can be potential productive uses of deepfakes [101], they may also imply negative economic effects [102], as well as severe legal ones[31].

Although an audit has revealed that the software to generate deepfake videos is still hard to use[32], there is an increase in such fake contents, not only affecting celebrities [33], but also less-known people[34,35]. As some authors have previously stated, technology will play a keystone role in fighting deepfakes [103]. In this sense, authors in [104] have very recently presented an approach to accurately detect fake portrait videos (97.29% accuracy) as well as to find out the particular generative model underlying a deep fake based on spatiotemporal patterns present in biological signals, under the assumption that a synthetic person, for instance, does not show a similar pattern of heart beat in comparison to the real one. Nevertheless, contributions are required from other fields such as legal, educational and political ones [101, 105].

As for some other open challenges related to cybersecurity, fake news and deepfakes require increasing the resources spent on detection technology; identification rates must be increased while the sophistication of disinformation continuously grows [102].

5.5. Final remarks

This work presents the results obtained from a comprehensive and systematic study of research papers, projects and initiatives concerning detecting fake news (online disinformation). Our goal was to show current and possible trends in this needed area of research in the computer science field due to the demands of societies from countries worldwide. Additionally, available resources (methods, datasets, etc.) to research in this topic have been thoroughly analysed.

In addition to the analysed previous work, the present study is aimed at motivating researchers to take up challenges in this domain, that increasingly impact current societies. More precisely, challenges still to be addressed are identified in order to propose exploring them.

Acknowledgement

This work is supported by the SocialTruth project[36], which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825477.

References

[1] S. B. Parikh, P. K. Atrey, Media-rich fake news detection: A survey, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018, pp. 436–441.
[2] X. Zhou, R. Zafarani, Fake news: A survey of research, detection methods, and opportunities, 2018. arXiv:1812.00315.
[3] N. K. Conroy, V. L. Rubin, Y. Chen, Automatic deception detection: Methods for finding fake news, Proceedings of the Association for Information Science and Technology 52 (2015) 1–4.
[4] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, Detection and resolution of rumours in social media: A survey, ACM Computing Surveys (CSUR) 51 (2018) 1–36.
[5] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–42.
[6] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, A. Krause, Fake news detection in social networks via crowd signals, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 517–524. doi:10.1145/3184558.3188722.
[7] J. Posetti, A. Matthews, A short guide to the history of'fake news' and disinformation, International Center for Journalists 7 (2018).
[8] E. H. Cline, 1177 BC: The year civilization collapsed, Princeton University Press, 2015.
[9] J. Neander, R. Marlin, Media and propaganda: The northcliffe press and the corpse factory story of world war i, Global Media Journal 3 (2010) 67.
[10] R. Herzstein, The most infamous propaganda campaign in history, GP Putnam & Sons (1978).

[31] https://spectrum.ieee.org/tech-talk/computing/software/what-are-deepfakes-how-are-they-created
[32] https://spectrum.ieee.org/tech-talk/computing/software/the-worlds-first-audit-of-deepfake-videos-and-tools-on-the-open-web
[33] https://www.bbc.com/news/av/technology-40598465
[34] https://www.bbc.co.uk/bbcthree/article/779c940c-c6c3-4d6b-9104-bef9459cc8bd
[35] https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them
[36] http://socialtruth.eu

[11] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management 57 (2020) 102025.
[12] H. Allcott, M. Gentzkow, Social Media and Fake News in the 2016 Election, Working Paper 23089, National Bureau of Economic Research, 2017. URL: http://www.nber.org/papers/w23089. doi:10.3386/w23089.
[13] S. Kula, M. Choraś, R. Kozik, P. Ksieniewicz, M. Woźniak, Sentiment analysis for fake news detection by means of neural networks, in: International Conference on Computational Science, Springer, 2020, pp. 653–666.
[14] M. de Cock Buning, A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation, Publications Office of the European Union, 2018.
[15] P. Canada. Parliament. House of Commons. Standing Committee on Access to Information, Ethics, B. Zimmer, Democracy under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly: Report of the Standing Committee on Access to Information, Privacy and Ethics, House of Commons= Chambre des communes, Canada, 2018.
[16] D. Collins, C. Efford, J. Elliot, P. Farrelly, S. Hart, J. Knight, G. Watling, Disinformation and" fake news': Final report, 2019.
[17] A. Giełczyk, R. Wawrzyniak, M. Choraś, Evaluation of the existing tools for fake news detection, in: IFIP International Conference on Computer Information Systems and Industrial Management, Springer, 2019, pp. 144–151.
[18] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge dataset, 2020. arXiv:2006.07397.
[19] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.
[20] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, M. Palomar, Fighting post-truth using natural language processing: A review and open challenges, Expert Systems with Applications 141 (2020). doi:10.1016/j.eswa.2019.112943.
[21] H. P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of research and development 1 (1957) 309–317.
[22] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of documentation (1972).
[23] N. Hassan, W. Gomaa, G. Khoriba, M. Haggag, Credibility detection in twitter using word n-gram analysis and supervised machine learning techniques, International Journal of Intelligent Engineering and Systems 13 (2020) 291–300. doi:10.22266/ijies2020.0229.27.
[24] A. Zubiaga, G. W. S. Hoi, M. Liakata, R. Procter, Pheme dataset of rumours and non-rumours, Figshare. Dataset (2016).
[25] P. Bharadwaj, Z. Shao, Fake news detection with semantic features and text mining, International Journal on Natural Language Computing (IJNLC) Vol 8 (2019).
[26] H. Wynne, Z. Wint, Content based fake news detection using n-gram models, 2019. doi:10.1145/3366030.3366116.
[27] S. Kaur, P. Kumar, P. Kumaraguru, Automating fake news detection system using multi-level voting model, Soft Computing 24 (2020) 9049–9069.
[28] P. Ksieniewicz, M. Choraś, R. Kozik, M. Woźniak, Machine learning methods for fake news classification, in: H. Yin, D. Camacho, P. Tiňo, A. J. Tallón-Ballesteros, R. Menezes, R. Allmendinger (Eds.), Intelligent Data Engineering and Automated Learning - IDEAL 2019 - 20th International Conference, Manchester, UK, November 14-16, 2019, Proceedings, Part II, volume 11872 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 332–339.
[29] B. D. Horne, S. Adali, This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, in: Eleventh International AAAI Conference on Web and Social Media, 2017.
[30] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, ACM, 2011, pp. 675–684.
[31] H. Telang, S. More, Y. Modi, L. Kurup, Anempirical analysis of classification models for detection of fake news articles, 2019. doi:10.1109/ICECCT.2019.8869504.
[32] S. Kong, L. Tan, K. Gan, N. Samsudin, Fake news detection using deep learning, 2020, pp. 102–107. doi:10.1109/ISCAIE47305.2020.9108841.
[33] L. Zhou, D. Zhang, Following linguistic footprints: Automatic deception detection in online communication, Communications of the ACM 51 (2008) 119–122.
[34] D. Zhang, L. Zhou, J. L. Kehoe, I. Y. Kilic, What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews, Journal of Management Information Systems 33 (2016) 456–481.
[35] A. Marouf, R. Ajwad, A. Ashrafi, Looking behind the mask: A framework for detecting character assassination via troll comments on social media using psycholinguistic tools, 2019. doi:10.1109/ECACE.2019.8679154.
[36] J. W. Pennebaker, M. E. Francis, R. J. Booth, Linguistic inquiry and word count: Liwc 2001, Mahway: Lawrence Erlbaum Associates 71 (2001) 2001.
[37] M. Ott, Y. Choi, C. Cardie, J. T. Hancock, Finding deceptive opinion spam by any stretch of the imagination, arXiv preprint arXiv:1107.4557 (2011).
[38] R. L. Robinson, R. Navea, W. Ickes, Predicting final course performance from students' written self-introductions, Journal of Language and Social Psychology 32 (2013) 469–479. doi:10.1177/0261927x13476869.
[39] C.-L. Huang, C. K. Chung, N. Hui, Y.-C. Lin, Y.-T. Seih, B. C. Lam, W.-C. Chen, M. H. Bond, J. W. Pennebaker, The development of the chinese linguistic inquiry and word count dictionary., Chinese Journal of Psychology (2012).
[40] M. del Pilar Salas-Zárate, E. López-López, R. Valencia-García, N. Aussenac-Gilles, Á. Almela, G. Alor-Hernández, A study on LIWC categories for opinion mining in spanish reviews, Journal of Information Science 40 (2014) 749–760. doi:10.1177/0165551514547842.
[41] D. I. H. Farías, R. Prati, F. Herrera, P. Rosso, Irony detection in twitter with imbalanced class distributions, Journal of Intelligent & Fuzzy Systems (2020) 1–17.
[42] B. Stoick, N. Snell, J. Straub, Fake news identification: A comparison of parts-of-speech and n-grams with neural networks, volume 10989, 2019. doi:10.1117/12.2521250.
[43] S. Dhuliawala, D. Kanojia, P. Bhattacharyya, Slangnet: A wordnet like resource for english slang, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 4329–4332.
[44] J. P. McCrae, I. Wood, A. Hicks, The colloquial wordnet: Extending princeton wordnet with neologisms, in: International Conference on

Language, Data and Knowledge, Springer, 2017, pp. 194–202.
[45] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., in: Lrec, volume 10, 2010, pp. 2200–2204.
[46] M. Thelwall, The heart and soul of the web? sentiment strength detection in the social web with sentistrength, in: Cyberemotions, Springer, 2017, pp. 119–134.
[47] A. A. Ahmed, I. Traore, Biometric recognition based on free-text keystroke dynamics, IEEE transactions on cybernetics 44 (2013) 458–472.
[48] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Security and Privacy 1 (2018) e9.
[49] A. Stolcke, J. Segal, Precise n-gram probabilities from stochastic context-free grammars, arXiv preprint cmp-lg/9405016 (1994).
[50] S. Kumar, K. Carley, Tree lstms with convolution units to predict stance and rumor veracity in social media conversations, 2020, pp. 5047–5058.
[51] K. S. Tai, R. Socher, C. D. Manning, Improved semantic representations from tree-structured long short-term memory networks, arXiv preprint arXiv:1503.00075 (2015).
[52] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations, arXiv preprint arXiv:1609.09028 (2016).
[53] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD Explor. Newsl. 19 (2017) 22–36. doi:10.1145/3137597.3137600.
[54] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management (2019).
[55] M. Chorań, A. Giełczyk, K. Demestichas, D. Puchalski, R. Kozik, Pattern recognition solutions for fake news detection, in: IFIP International Conference on Computer Information Systems and Industrial Management, Springer, 2018, pp. 130–139.
[56] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, Commun. ACM 59 (2016) 96–104. doi:10.1145/2818717.
[57] S. Afroz, M. Brennan, R. Greenstadt, Detecting hoaxes, frauds, and deception in writing style online, in: Proceedings of the 2012 IEEE Symposium on Security and Privacy, SP '12, IEEE Computer Society, 2012, pp. 461–475. doi:10.1109/SP.2012.34.
[58] Z. Jin, J. Cao, Y. Zhang, J. Zhou, Q. Tian, Novel visual and statistical image features for microblogs news verification, Trans. Multi. 19 (2017) 598–608. doi:10.1109/TMM.2016.2617078.
[59] C. Chen, K. Wu, S. Venkatesh, X. Zhang, Battling the internet water army: Detection of hidden paid posters, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013) (2011) 116–120.
[60] G. Gravanis, A. Vakali, K. Diamantaras, P. Karadais, Behind the cues: A benchmarking study for fake news detection, Expert Systems with Applications 128 (2019) 201 – 213.
[61] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences 497 (2019) 38 – 55.
[62] C.-S. Atodiresei, A. Tănăselea, A. Iftene, Identifying fake news and fake users on twitter, Procedia Computer Science 126 (2018) 451 – 461. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.
[63] J. Buford, H. Yu, E. K. Lua, P2P networking and applications, Morgan Kaufmann, 2009.
[64] K. Xu, F. Wang, H. Wang, B. Yang, Detecting fake news over online social media via domain reputations and content understanding, Tsinghua Science and Technology 25 (2019) 20–27.
[65] R. Hegli, H. Lonas, C. K. Harris, System and method for developing a risk profile for an internet service, 2013. US Patent 8,438,386.
[66] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, N. Feamster, Building a dynamic reputation system for dns., in: USENIX security symposium, 2010, pp. 273–290.
[67] P. Lison, V. Mavroeidis, Neural reputation models learned from passive dns data, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 3662–3671.
[68] B. Rahbarinia, R. Perdisci, M. Antonakakis, Segugio: Efficient behavior-based tracking of malware-control domains in large isp networks, in: 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, IEEE, 2015, pp. 403–414.
[69] Y. Tang, S. Krasser, P. Judge, Y.-Q. Zhang, Fast and effective spam sender detection with granular svm on highly imbalanced mail server behavior data, in: 2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing, IEEE, 2006, pp. 1–6.
[70] K. Shu, H. R. Bernard, H. Liu, Studying fake news via network analysis: detection and mitigation, in: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, Springer, 2019, pp. 43–65.
[71] X. Zhou, R. Zafarani, Network-based fake news detection: A pattern-driven approach, ACM SIGKDD Explorations Newsletter 21 (2019) 48–60.
[72] K. Shu, S. Wang, H. Liu, Exploiting tri-relationship for fake news detection, arXiv preprint arXiv:1712.07709 8 (2017).
[73] L. Bondi, S. Lameri, D. Güera, P. Bestagini, E. J. Delp, S. Tubaro, Tampering detection and localization through clustering of camera-based cnn features, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017, pp. 1855–1864.
[74] R. Zhang, J. Ni, A dense u-net with cross-layer intersection for detection and localization of image forgery, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 2982–2986.
[75] A. James, E. B. Edwin, M. Anjana, A. M. Abraham, H. Johnson, Image forgery detection on cloud, in: 2019 2nd International Conference on Signal Processing and Communication (ICSPC), IEEE, 2019, pp. 94–98.
[76] N. Kanwal, A. Girdhar, L. Kaur, J. S. Bhullar, Digital image splicing detection technique using optimal threshold based local ternary pattern, Multimedia Tools and Applications (2020) 1–18.
[77] S. Dua, J. Singh, H. Parthasarathy, Detection and localization of forgery using statistics of dct and fourier components, Signal Processing: Image Communication (2020) 115778.
[78] A. K. Jaiswal, R. Srivastava, A technique for image splicing detection using hybrid feature set, Multimedia Tools and Applications (2020) 1–24.
[79] J. N. Jothi, S. Letitia, Tampering detection using hybrid local and global features in wavelet-transformed space with digital images, Soft Computing 24 (2020) 5427–5443.

[80] M. Bilal, H. A. Habib, Z. Mehmood, T. Saba, M. Rashid, Single and multiple copy–move forgery detection and localization in digital images based on the sparsely encoded distinctive features and dbscan clustering, Arabian Journal for Science and Engineering (2019) 1–18.
[81] P. Suryawanshi, P. Padiya, V. Mane, Detection of contrast enhancement forgery in previously and post compressed jpeg images, in: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), IEEE, 2019, pp. 1–4.
[82] Y. Guo, X. Cao, W. Zhang, R. Wang, Fake colorized image detection, IEEE Transactions on Information Forensics and Security 13 (2018) 1932–1944.
[83] M. Choraś, M. Pawlicki, R. Kozik, K. Demestichas, P. Kosmides, M. Gupta, Socialtruth project approach to online disinformation (fake news) detection and mitigation, in: Proceedings of the 14th International Conference on Availability, Reliability and Security, 2019, pp. 1–10.
[84] J. Dong, W. Wang, T. Tan, Casia image tampering detection evaluation database, in: 2013 IEEE China Summit and International Conference on Signal and Information Processing, IEEE, 2013, pp. 422–426.
[85] Y. Zheng, Y. Cao, C.-H. Chang, A puf-based data-device hash for tampered image detection and source camera identification, IEEE Transactions on Information Forensics and Security 15 (2019) 620–634.
[86] V. Christlein, C. Riess, J. Jordan, C. Riess, E. Angelopoulou, An evaluation of popular copy-move forgery detection approaches, IEEE Transactions on information forensics and security 7 (2012) 1841–1854.
[87] E. Ardizzone, A. Bruno, G. Mazzola, Copy–move forgery detection by matching triangles of keypoints, IEEE Transactions on Information Forensics and Security 10 (2015) 2084–2094.
[88] M. Castro, D. M. Ballesteros, D. Renza, A dataset of 1050-tampered color and grayscale images (cg-1050), Data in brief 28 (2020) 104864.
[89] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).
[90] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations., in: ICWSM, 2015, pp. 258–267.
[91] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (2020) 171–188.
[92] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.
[93] B. Krawczyk, L. Minku, J. Gama, J. Stefanowski, M. Wozniak, Ensemble learning for data stream analysis: A survey, Information Fusion 37 (2017) 132–156. doi:10.1016/j.inffus.2017.02.004.
[94] S. Wang, L. L. Minku, X. Yao, Resampling-based ensemble methods for online class imbalance learning, IEEE Trans. Knowl. Data Eng. 27 (2015) 1356–1368.
[95] P. Ksieniewicz, Z. Paweł, C. Michał, K. Rafał, G. Agata, W. Michał, Fake news detection from data streams, in: Proceedings of the International Join Conference on Neural Networks, 2020.
[96] Z. Chen, B. Liu, R. Brachman, P. Stone, F. Rossi, Lifelong Machine Learning, 2nd ed., Morgan & Claypool Publishers, 2018.
[97] A. Pentina, C. H. Lampert, Lifelong learning with non-i.i.d. tasks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, p. 1540–1548.
[98] Yaochu, Jin, et al., Neural network regularization and ensembling using multi-objective evolutionary algorithms, in: CEC'04), volume 1, 2004, pp. 1–8. doi:10.1109/CEC.2004.1330830.
[99] M. Choraś, M. Pawlicki, D. Puchalski, R. Kozik, Machine learning–the results are not the only thing that matters! what about security, explainability and fairness?, in: International Conference on Computational Science, Springer, 2020, pp. 615–628.
[100] R. Chawla, Deepfakes: How a pervert shook the world, International Journal of Advance Research and Development 4 (2019) 4–8.
[101] M. Westerlund, The emergence of deepfake technology: A review, Technology Innovation Management Review 9 (2019).
[102] A. O. J. Kwok, S. G. M. Koh, Deepfake: a social construction of technology perspective, Current Issues in Tourism 0 (2020) 1–5. doi:10.1080/13683500.2020.1738357.
[103] M.-H. Maras, A. Alexandrou, Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos, The International Journal of Evidence & Proof 23 (2019) 255–262. doi:10.1177/1365712718807226.
[104] U. A. Ciftci, I. Demir, L. Yin, How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals, 2020. arXiv:2008.11363.
[105] J. Pitt, Deepfake videos and ddos attacks (deliberate denial of satire) [editorial], IEEE Technology and Society Magazine 38 (2019) 5–8.