UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Data Science

Profiling Fake News Spreaders: Personality and Visual Information Matter

Relatrice: Prof.ssa Gabriella Pasi
Correlatore: Prof. Paolo Rosso

Tesi di Laurea Magistrale di:
Riccardo Cervero
Matricola 794126

Anno Accademico 2019-2020

Abstract

In the post-truth era, users are overwhelmed by massive and continuous flows of information, within which it is often difficult to distinguish credible content from those that intentionally or erroneously amplify the effects of disinformation. In this context, fake news can spread by exploiting the cognitive bias of readers, especially in virtual ecosystems without the intermediation of domain experts who can filter the correct messages. Malicious algorithms - like bots - are therefore trained to learn lexical and semantic patterns aimed at provoking specific emotional processes, such as anger and disgust, and persuading human consumers. To obstruct this phenomenon, psychological and social factors underlying this vulnerability must therefore be examined. As a first objective, this thesis thus proposes to analyze the correlation between psycho-linguistic patterns extractable from user's posts and the tendency to spread false information, through Five Factor and LIWC models. Since online contents often exploit multimedia elements to attract the audience, a methodology aimed at profiling the authors based on the images they incorporated in the tweets has then been tested, to verify whether different visual patterns characterize 'fake news spreader' and 'real news spreaders'. Furthermore, the effectiveness of these features sets will be evaluated based on their ability to improve the results of state-of-the-art methods like the best performing models at the Author Profiling Task at PAN 2020.

Acknowledgments

I would like to thank Professor Gabriella Pasi very much for passionately guiding me in my choice of thesis topic and for helping me to complete the final work.

I would also like to express my deep gratitude to Professor Paolo Rosso for training me during the traineeship activity described in this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent developments in Information Technologies have made it possible to convey a huge amount of information with very low latency and an increasing ease of use, even for less educated users. This trend of democratization in production of online content has obvious advantages, offering anyone able to access the platforms the opportunity to reach many other users, for example for scientific or reporting purposes. However, precisely because of this more fluid circulation of contents, it is evident how an irresponsible use of this open systems can cause, in the absence of strict controls, serious damages to the virtual community itself. One of the main consequences consists in an uncontrolled diffusion of false information, whose manufacture has evolved during centuries into the present powerful and sophisticated forms. Efforts to counter this dangerous dynamic are based primarily on the definition of the types of false information, as it was done in [1] and [2]. First of all, the intention to cause harm is commonly regarded as the main criterion: thus, the two sub-phenomena of misinformation and disinformation are distinguished.

1.1 Categories: misinformation and disinformation

1.1.1 Misinformation

Misinformation is characterized by being devoid of the intent to damage, and includes categories such as unverified or partially verified stories, more often due to errors of journalists and media in reporting news, or to the re-sharing of a wrong interpretation of a news event by non-expert readers. Either way, these fake or half-true content pass from one channel to another without serious check of the original and secondary sources, and even of the content itself.

A particular type of misinformation is then satire, which, through a figurative language and rhetorical devices - irony and sarcasm -, offers a humorous narration aimed at ridiculing or criticizing people mainly belonging

to the contemporary political environment. Although satirical news deliberately feature hilarious or surreal stories, it's not that uncommon for them to be misinterpreted as realistic and propagated as such.

However, despite of being not directly aimed at damaging the quality of information circulating in the virtual environment, all these dynamics are potentially capable of generating a serious side effect: erroneously modifying readers' opinion and, in some cases, radicalising it. This can be defined as the misinformation effect: information presented after a given event interferes with the retention of previously stored information and distort the original memory [3]. Loftus, in [4, 5], highlights how the extent of misinformation effect depends on two main causes: misattribution - i.e. the error in the attribution of the source - and suggestibility, which corresponds to the influence of expectations. The susceptibility is in turn linked to some receiver's characteristics, such as imagery abilities, age, neurological conditions and, most importantly, personality traits. A significant contribution from this point of view is the one offered by Briggs-Myers Myers [6]: after having subjected a sample to the same experiments previously conducted by Loftus, a model - Myers Briggs Type Indicator (MBTI) - was generated that assesses the incidence of the misinformation effect on a given subject based on his personality characteristics. It has thus been observed, for example, that introversion is more often associated with a lower confidence in one's own memory and a greater tendency to accept inaccurate information [7, 8]. It has also been shown that a greater reception of misinformation contents depends on the level of empathy, engagement and self-monitoring propensities [4].

In addition to these observations, Zajonc [9] has also demonstrated that the mere exposure - especially if excessive - to an idea or message generates a positive attitude towards it. From this, we can deduced that becoming part of a community that promotes a certain narrative, albeit a toxic one, can profoundly influence an individual's belief system.

1.1.2 Disinformation

Impressing the reader with the aim of persuading him becomes, instead, a premeditated strategy on the part of those who produce and share contents belonging to the category of disinformation, or more commonly defined as 'fake news'. Fake news can, in fact, be described as false articles intentionally fabricated to mislead the audience [1]. Above all, lately, thanks to the amount of data shared knowingly or unknowingly by users on the Web, it has become possible a 'custom' fabrication of these artifact content, resulting in what can be called unfair marketing.

The categories proposed in the following paragraphs may often overlap.

Clickbaiting The least dangerous case occurs in the context of the so-called 'yellow journalism' [10]: this practice is used by the media and news organizations to attract more online readers and increase online traffic. This is the case with 'clickbaiting', referring to the inclusion in the article of titles and descriptions prone to exaggeration and sensationalization. In details, Chen et al. [11] proved that the attractiveness of such posts is based on the recurring use of a reversal narrative style with forward referencing - in order to tease reader's curiosity and engage him emotionally -, image insertion, and above all linguistic patterns and topics adapted to consumers' behaviors and personality traits. Even though this type of content often presents a strong inconsistency between title or images inserted and the actual content of the article - and it is, therefore, not very credible -, a report by the website PolitiFact [12] analyzed the growth of this phenomenon, revealing even the existence of a market of clickbaiting contents fabricated for specific users and distributed on many web pages.

Hoaxes Another popular type of fake news within social networks is the hoax, a story that reports a distorted and inaccurate description of a news item, presenting it as a verified truth [13]. Sometimes, these narratives just offer an opposite view with respect to the public opinion, for example fueling irrational concerns about positive news. In some cases, they attribute false citations to known people, as happened with the spread of false articles about the endorsement of Pope Francis for Donald Trump's presidency in 2016[1]. If more sophisticated, they mix fact and fiction to narrate events that never really happened, such as the discovery of an extraterrestrial civilization on the opposite side of the moon, alleged by a series of YouTube videos which provided as evidence modified versions of official photos released by NASA[2]. To give other examples, fake posts are often published reporting the death of a celebrity. Regarding the so-called 'death hoaxes' sub-category, Situngkir [14] analyzes, from an experiment on Indonesian Twitter accounts, the rapid expansion of this contents, even if originally shared by profiles with a small number of followers. In addition, his results - upholding the conclusions of Oh et al. [15] - indicate that not only official medias, but also popular accounts, can filter hoaxes and brake their dissemination.

Rumors A similar class, but more difficult to identify and debunk is that of rumors, consisting in stories whose credibility appears ambiguous or still to be evaluated. The second main characteristic is the spontaneity with whom these type of fake news are generated and developed, without necessarily

[1]Website Snopes debunked this fake news: https://www.snopes.com/fact-check/pope-francis-donald-trump-endorsement/.

[2]The 'Apollo 20 hoax': https://www.ibtimes.co.uk/world-ufo-day-alien-girl-secret-apollo-20-mission-1454928.

having to rewrite an event. This expansion is, sometimes, so uncontrolled that it foster a vast public debate and influences political dynamics, as the allegations stating that Barack Obama secretly practices religion of Islam [16]. It is interesting to note, as illustrated in [17], that the growth of true rumors tends to extinguish faster than that of false rumors. More over, empirical test [18] reveal the existence of a particular recurring writing style in online rumors. Rumors are, by the way, an ancient phenomenon: an example is the Shakespeare's authorship question[3], circulated since the middle of the 19th century.

Deceiving contents Ultimately, though, the most dangerous category of fake news includes multimedia contents built not only to convince of their own truthfulness, as happens with hoaxes and rumors, but properly with the goal of manipulating the opinion of users and deeply modifying their consumption habits [19], guide their political ideologies or undermine their trust in institutions. The vast potential of this increasingly debated phenomenon is nowadays well-known, able, for example, to deceive non-expert readers about complex scientific issues. A famous and extremely current example can be found in the huge wave of disinformation related to the Coronavirus pandemic, which is referred to by the new term 'infodemic' [20]: the suggestion to consume alleged 'alternative' treatments, such as the anti-malarial drug called hydroxychloroquine, which has even been recommended by Donald Trump[4] even though it has been shown not only to be ineffective, but also to have potentially fatal complications in combination with COVID-19 disease[5]; the unfounded allegations - also promoted by the Nobel Prize winner Montagnier - that the virus did not originate naturally but was created in a laboratory in the city of Wuhan and intentionally spread[6]; the distrust of vaccines efficacy[7].

Another well-known case of mistrust of the scientific view is the denialism of human-caused global warming, strongly promoted by the oil industry[8] and supported by certain politicians such as Donald Trump [21] and Brazilian President Jair Bolsonaro [22]. In these circumstances, it is reasonable

[3]Shakespeare authorship question: wikipedia.com/Shakespeare_authorship_question

[4]Trump's statements about hydroxychloroquine: https://www.cnbc.com/trump-hydroxychloroquine.html.

[5]Some studies about the ineffectiveness and the side-effects of hydroxychloroquine are reported here: https://www.bbc.com/news/51980731 and https://www.theguardian.com/hydroxychloroquine-trumps-covid-19-cure-increases-deaths-global-study-finds.

[6]Montagnier's statements and fact checking articles are available here: https://www.connexionfrance.com/French-news/Disputed-French-Nobel-winner-Luc-Montagnier-says-Covid-19-was-made-in-a-lab-laboratory.

[7]Most popular fake news about vaccines: https://www.washingtonpost.com/faq-coronavirus-vaccine-misinformation/.

[8]An analysis of the relationship between the oil industry and climate denialism: https://www.bbc.com/news/stories-53640382.

to assume that behind the publication of these contents there is often the need to perpetrate a given economic interest. In addition, false articles are manufactured to unfairly harm economic competitors or directly manipulate the financial market, as examined by Kogan et al. in [23]. Indeed, in 2017 U.S. Securities and Exchange Commission filed charges[9] against several public companies, communications businesses, and hundreds of individuals after discovering what have been defined a 'sophisticated stock promotion system'. The mechanism of this scheme was to spread false news about listed companies in order to increase their shares values in the stock exchanges. One case is that of ImmunoCellular Therapeutics', whose share price increased by 263% in one months after the publication of an article claiming that the company had released an highly effective treatment for cancer[10].

Propaganda However, recently, the most debated consequences by public opinion concern the vulnerabilities revealed by political and electoral dynamics when disturbed by the dissemination of non-credible information. On this subject, Zannettou et al. [2] gather under the term propaganda all those contents that form part of a wider communication strategy aimed at manipulating the awareness and interpretation of receivers about certain topics or fact belonging the contemporary political context. This strategy may aim at retaining political supporters, damaging detractors, or even, in extreme cases, suppressing rebellions. Indeed, a characteristic highlighted by Zannettou et al. [2] is the ability to profoundly influence not only the electoral outcome, but various aspects of the evolution of the societies with which the phenomenon interferes.

The influence of the disinformation industry within the political sphere has become a key issue for national security and for the defence of the democratic social structure, especially after the scandal that saw the private company Cambridge Analytica [24, 25] steal data from many millions of Facebook profiles[11] and reuse it to produce tailor-made content that affects users and manages to engage them emotionally enough to persuade certain voting choices. In details, personal information was collected without any consent by a mobile application which - in addition to storing sensitive data such as location, likes, messages, timeline posts and even friends - proposed questions aimed at estimating a psychographic profile of users. This psychological description was detailed enough to effectively indicate the best type of advertisement for a given profile category, in a specific location, for a precise political event. There is evidence that this system was adopted by Republican

[9] Press release on SEC investigations: https://www.sec.gov/news/press-release/2017.

[10] Fake news infiltration into financial market: https://www.ft.com/content/a37e4874-2c2a-11e7-bc4b-5528796fe35c.

[11] Although Cambridge Analytica claimed that 30 million profiles were analysed, Facebook later stated that the data actually belonged to over 87 million users, of which approximately 81% from the U.S. [26].

candidate Ted Cruz during the presidential campaign in 2016. An evolution of this system was apparently implemented during Donald Trump's presidential campaign in the same year to retain supporters and attract undecided voters. In this case, after analyzing the interactions between users and content, two types of posts were generated: Trump supporters were offered a glowing narrative about the candidate and instructions about polling locations; so-called 'swing voters' received negative messages about the work of the opponent, Hilary Clinton [27]. This technique was also allegedly used to divert voting decisions during the United Kingdom's referendum about leaving the European Union in 2016: according to an ongoing investigation by the UK Electoral Commission, there probably were strong interferences by Russian-sourced accounts in the referendum debate, and it is assumed that Kremlin's support for the 'Leave' vote was behind these suspicious activities. These virtual intrusion consisted in the production and dissemination of fake news, especially on Twitter, where it is estimated that hundreds of thousands of Russian bots and trolls shared millions of tweets and were then deleted the day after the outcome [28, 29, 30, 31, 32, 33]. Similar strategies seem to have been adopted in several other countries all over the world: other Russian fake Twitter profiles tried to influence the Scottish independence referendum in 2014 [34]; Cambridge Analytica organised a disinformation campaign in favour of the Maltese Labour Party in the 2013 election [35]; the Italian party Lega Nord has increased its popularity thanks to the study of profiling strategies for a more effective communication on social networks [36]; more than 500000 users were targeted by false articles in favour of the Indian National Congress during Indian elections in 2010, 2018 and 2019 [37, 38]; the same micro-targeting methodology has been exploited for Uhuru Kenyatta's presidential campaign in Kenya in 2013 and 2017 [39]; some evidence points to similar collaborations between Cambridge Analytica and the Institutional Revolutionary Party in Mexico [40], as well as between the private company and Rodrigo Duerte's party in Philippines [41].

Such a widespread phenomenon[12] has provoked a regulatory clampdown by governments and international organisations. In particular, in the United States, Facebook was fined for allowing such serious violations of users' privacy. The European Union launched the General Data Protection Regulation [42], which regulates the stages of sensitive and personal data processing by all entities offering services to EU citizens, including non-EU companies.

Finally, it is worth mentioning a special subclass of propaganda: the so-called "firehose of falsehood" - also "firehosing" -, which defines a huge and continuous flow of messages biased towards an idea, spread consistently over multiple channels [43, 44]. The strong persuasiveness of this typology is based precisely on the involvement of a wide variety of sources, because,

[12] In 2018, Cambridge Analytica's CEO declared that the company had been involved in more than 200 elections around the world.

as studies have shown [45, 46], people expect that information from multiple sources is likely to reflect different perspectives and is therefore worth greater consideration. Further research [47, 48] explains that when the interest about a certain topic is low, the effectiveness of a message can depend more on the number of arguments supporting it than on the quality of those arguments. The second characteristic of volume is also crucial for the success of the manipulation attempts: if the amount of posts supporting an idea - even if it is wrong - is sufficient, then readers tend to prefer their peers' view instead of debunking articles by experts.

It is assumed that this technique was effectively used by Russian accounts during the tensions arisen from the annexation of Crimea since 2014.

Conspiracy Theories The increasing speed at which false information is generated and disseminated is also due to the new consumers' tendency to rely more on unofficial media to keep themselves informed. A survey conducted by Pew Research Center [49] disclosed in 2018 that 68% of U.S. adults receive their news also through social networks, and that a third of these do so frequently. In particular, according to another study by Pew Research Center [50], this trend is peculiar to the younger age group: in the 18-29 year-old range, 36% keep themselves informed through social networks, while 27% of respondents stated they prefer news websites, 16% television. This trend has grown sharply: in 2020, the percentage of young people who say they rely only on social media rose to 48%; the same happened in the range 30-49 years, as observed in [51]. The problematic aspect is that, although most of the content published on these platforms comes from unofficial sources, only just over half of the adults surveyed (57%) are sceptical about the accuracy of the content they receive on social media. Furthermore, the survey shows that almost half (48%) believe that receiving news from social networks does not affect their ability to understand current events, and that 36% are convinced that it even offers them a better understanding with respect to the other media; only 15% thus say that news spread on social networks make them more disoriented.

It also emerged that the feature that most encourages people to use social media as a source of information is, along with convenience, the possibility of being able to interact with other users about the discussed topics. However, as opposed to the one-sided communication between traditional media and consumers, these strong peer-to-peer connections fostered by social platforms in the Web 2.0 risk making users even more vulnerable to disinformation and misinformation, and more susceptible to polarized interpretations of current events. This is what emerges from a third investigation by Pew Research Center [51] in 2020, who studied the correlation between the news consumption habits of U.S. adults and their way of perceiving and interpreting events. Firstly, it was observed that a greater use of social media as a primary news

source is associated with a less accurate knowledge and awareness about the topics: after asking respondents a series of 29 questions related to popular current events - such as about the Coronavirus pandemic, about Donald Trump's impeachment trial and about economic reforms enacted in the same year - the lowest number of correct answers was recorded for the group of respondents who say they only get their information through social media. This is particularly serious when dealing with issues related to the political context. In fact, a knowledge index based on the number of correct answers on the political topics was considered: 'high' if greater than 89% of the total number of questions, 'middle' if between 56% and 88%, and 'low' if less than 55%. The result is visible in Figure 1-I: within the group of respondents who say they keep themselves informed through social media, there is a large percentage of individuals with low political knowledge (57%) and a very small proportion with high knowledge (17%). This category, therefore, records the second worst result after those who receive news from local television. The media associated with a deeper understanding of the current political context are news website and radio, both of which show more than double the percentage of individuals with high familiarity with political issues with respect to the social media group (respectively 45% and 42%).

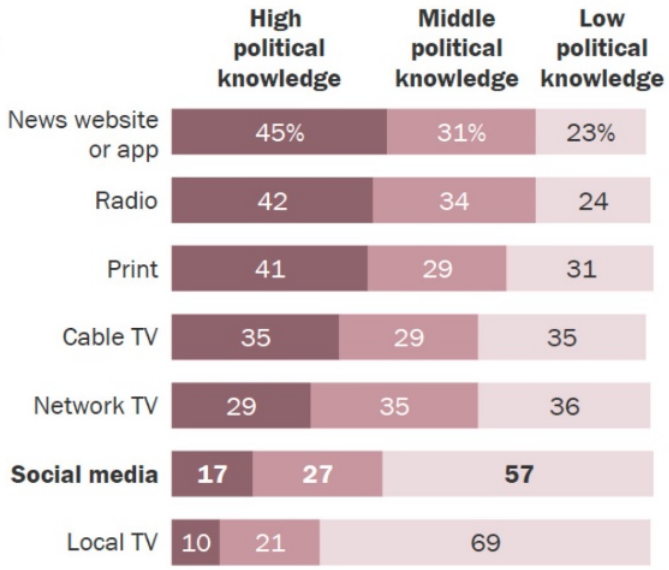

Figure 1-I: The figure shows the percentage of respondents - divided by the category of media they most rely on to keep themselves informed - who show respectively high, medium or low knowledge about the proposed political topics.
Source: '*Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable*', Pew Research Center (2020) [51].

Above all, according to this survey, this lowered awareness about the debated issues exposes those individuals who most often inform themselves on social media to a particular category of disinformation: conspiracy theories. Conspiracy theories are paranoid narratives alternative with respect to the prevailing opinion, which interpret specific events as the result of secret plots hatched by sinister and powerful parties, even when more convincing explanations are available [52, 53, 54, 55, 63]. A recently popular example is the so-called 'QAnon' phenomenon [57]: a conspiracy theory linked to extreme right-wing milieus - like the so-called 'Alternative Right' or simply 'Alt-Right' [58] -, according to which a 'deep state' perpetrated a secret plot against former U.S. President Donald Trump and his supporters. The genesis of the movement can be attributed to the first post by an anonymous user[13] on the 4chan platform[14], which falsely claimed to have access to information proving the involvement of Trump's detractors - including artists, liberal journalists, Democratic politicians and high-ranking officials[15] - in international child trafficking, satanic worship and cannibalism. The post also accused members of this secret sect of attempting to overthrow Trump's presidency in a coup d'état. The narrations promoted by the movement have been widely diffused and discussed around the world. For the first time in history, a conspiracy theory is officially regarded as a risk to internal security: in August 2019, the F.B.I. declared the QAnon phenomenon to be a potential source of domestic terrorism[16]. Despite this, the movement has been welcomed by the U.S. government's top brass and some of its promoters have been invited to the White House by Donald Trump[17]. The rapid polarization of the movement meant that members not only spread false allegations, but also promoted anti-democratic and violent ideologies. This escalation culminated on 6 January 2021, when the followers of this conspiracy theory organised and executed - with other similar Alternative Right groups[18] - a pro-Trump mob during the counting of the votes of 2020 U.S. presidential election on Capitol Hill [60, 61]. The assault was regarded as unprecedented in the history of American politics, and resulted in five deaths and 52 arrests. In many other cases, this type of online fake news has

[13]The account was named 'Q'. Hence, the name 'QAnon' derived from the combination with the abbreviation 'Anon'.

[14]4chan is an imageboard website originally created for image publishing and discussion about manga and anime, but often associated with various internet subcultures, notably the Alt-Right, Anonymous [59] and its Chanology project.

[15]The leaders of the alleged deep state would be Barack Obama, Hillary Clinton and George Soros.

[16]QAnon movement declared source of domestic terrorism by F.B.I, August 2019: https://www.washingtonpost.com/fbi-warning-about-qanon.

[17]Donald Trump often praised QAnon: https:/www.buzzfeednews.com/trump-praises-qanon.

[18]Among the other Alternative Right groups identified during the assault to Capitol Hill there were 'Proud Boys', 'Traditionalist Worker Party', 'Three Percenters', 'Groyper Army', Holocaust deniers, etc.

caused such an hysteria that it has resulted in serious attacks and crimes. A well-known and recent case is the one related to the 'Pizzagate' conspiracy theory [62], from which the QAnon narrative derives: a man, convinced that some Democratic politicians were perpetrating paedophile acts with the complicity of some restaurant chains, went to one of these restaurants and fired an assault rifle at its owner.

In general, regardless of the topic discussed, conspiracy theories are a category of disinformation that is very difficult to counteract: their followers regard refutation as further proof of the truth of the theory, almost turning the act of believing into a matter of faith [63]. Studies have also shown a correlation between the tendency to believe in such theories and specific personality traits such as paranoia, insensitivity and indifference to ethics [64]. In addition, countering the formation of these disinformation bubbles becomes even more difficult in the absence of a rapid and effective intermediation by experts who can refute false information, especially when the social networks on which they circulate adopt content moderation rules that are not strict enough to prevent the spread of dangerous messages. As a matter of fact, recently we are witnessing the 'migration' of these user communities to platforms that do not operate any serious control on posts, by virtue of a supposed 'freedom of expression'. Proof of this is the huge growth in the popularity of Reddit[19], which, despite all its controversies [65, 66], in 2020 stands as the thirteenth most used social network in the world[20], with 430 million monthly active users, surpassing even Twitter (in seventeenth position with 330 million monthly active accounts). Another emblematic case concerns the microblogging service Parler[21], which presented itself as a truly impartial alternative to mainstream social platforms such as Twitter and Facebook, making explicit its intention not to restrict any kind of content. Parler became popular especially in the American right-wing circles and among Trump supporters, doubling its subscribers and increasing its number of active users eightfold only during the last phase of the 2020 U.S. presidential campaign [67]. After it was discovered that the assault to Capitol Hill on 6 January 2021 was coordinated via Parler, the mobile application has been removed and the platform has been erased from Amazon's servers, going offline[22].

[19] Official Reddit website: https://www.reddit.com.

[20] The ranking of social platforms with at least 100 million monthly active users is available at: https://en.wikipedia.org/Social_platforms_at_least_100_million_active_users.

[21] Official Parler website: https://parler.com.

[22] Amazon, Apple and Google Cut Off Parler: https:/www.nytimes.com/apple-google-parler.

1.1.3 Disinformation Motivations

Given this classification, taking inspiration from both [1] and [2], it is possible to distinguish some main motivations for spreading fake news.

Firstly, the purpose is, in most cases, the quest for better profits or the achievement of greater popularity: viral contents attract the online users' traffic and thus produces higher advertising revenue [12, 68].

Less frequently, the objective may be ideological: the false information is in this case aimed, for example, at advantaging political candidates, damaging opponents or generating a debate [69]. Similar reasons consist in the manipulation of the public debate concerning a particular issue, the influence of the readers' behaviours - e.g. consumption -, or the damage of the public image related to specific people, organisations or entities.

In other instances, organisations or civilians share information to create confusion among readers on a topic, e.g. with the aim of sowing discord in a foreign country [70].

One factor that often encourages fake news dissemination, especially on social networks platforms, is fun: this is the case with users called 'trolls', whose aim is purely personal amusement [71].

Finally, a false narrative can propagate simply because a considerable number of users are passionate about it and their judgement is penalised, making them unable to filter out the correct information [72].

1.2 Project goals

On the basis of what has been said in the previous Chapter 2.3, it is therefore reasonable to assume that the phenomenon of disinformation does not depend solely on the persuasive capacity of the single fabricated news items, but rather on the psychological and cognitive predisposition of the non-expert user, and his own intellectual weaknesses. From this follows, for example, the paradoxical popularity of theories that are not credible at all, and the great efforts with which they have to be refuted. At this juncture, in order to counter the propagation of fake news, it is straightforward the need to identify which users are most inclined to create or share fake news by learning the cognitive and psychological vulnerabilities that are most often associated with this lower capacity of discerning real news. Therefore, in this project we will tackle the task of users' profiling, with the specific aim of classifying them into two classes:

- fake news spreaders, defined as individuals who show a tendency to create and/or share false information on the Web;
- real news spreaders, those who never publish news that is verified as false.

Assuming that the aforementioned vulnerabilities may derive from certain psychological inclinations, the first aim is to demonstrate that the 'fake news spreaders' are associated with a different set of personality traits compared to readers with the ability to filter realistic content. Thus, as a first part of this project, methodologies will be proposed capable of extracting personality characteristics from the text shared by a sample of users, in order to subsequently evaluate its impact on the tendency to spread false content and, finally, to test, in combination with various predictive models, their effectiveness for the task of binary classification into 'fake news spreader' and 'fake news non-spreader'.

As it will be explained in Chapter 2 (Section 2.3.2), it is then speculated that fake news spreaders - whether human or automated, like bot - publish images with precise characteristics. The second objective of the work, therefore, will be to investigate this association between visual patterns and the tendency to spread fake news.

The contribution of the project undertaken during the thesis period can be summarised in the following three points:

1. Inspired by the CheckerOrSpreader model by Giachanou et al. [122], we have defined and evaluated the effectiveness of a Machine Learning based approach that exploits both personality and psycho-linguistic features respectively extracted through the Five Factor Model and the LIWC model - in combination with other emotional dimensions and Bag-Of-Words (BOW) features - to binary classify users into 'fake news spreader' and 'fake news non-spreader'

2. Inspired by [123], we have also analysed the effectiveness of visual features extracted by pre-trained convolutional neural networks - alone or mixed with the sets of personality information - to perform a binary classification of users into real and fake news spreaders;

3. We have verified the feasibility of improving the effectiveness of state-of-the-art approaches by incorporating and/or replacing the proposed personality and visual information into the best performing solutions presented at the Author Profiling Task at PAN 2020[23].

The thesis manuscript is organised into 8 Chapters, shortly described in the following. Chapter 3 presents the theoretical framework and the related works about the task of fake news detection in social media and its state of the art. In particular, all categories of approaches used to tackle this task

[23]Author Profiling task at PAN 2020: https://pan.webis.de/clef20/pan20-web/author-profiling.html.

will be illustrated, divided into the two macro-categories of *knowledge-based* and *data-driven*. In particular, Section 3.2 introduces the background of the Author Profiling perspective, in particular what was done in the shared task at PAN 2020: the dataset provided, a general description of the submitted solutions and the best performances. Chapter 4 discusses in detail the best performing models at the Author Profiling task at PAN 2020: Buda-Bolonyai's and Pizarro's systems. In addition, the CheckerOrSpreader architecture by Giachanou et al. will be introduced in this Chapter. The next two Chapters (5 and 6) describe the main methodologies behind the contribution of the thesis work: Chapter 5 illustrates the two methods of extraction of personality and psycho-linguistic features starting from the plain text of the tweets, respectively with the LIWC resource and the Five Factor Model; Chapter 6 deals with explaining the collection of images from each user's sub-corpus and how visual information is finally obtained. Lastly, Chapter 7 describes all the experiments carried out to answer all the research questions indirectly posed by each project objective. This Chapter will present all the variations realized to the trained models, all the tested combinations among the features and the obtained results, which will then be summarized and commented in the last Chapter 8.

Chapter 2

Fake News Spreading

Although all types of fake news described in the previous Chapter can be debunked by the explanations of domain experts in a potentially easy way, a range of factors makes this activity extremely complex, preventing the proper assessment of the credibility of information in the Web: these include *(i)* the growing use of social networks as a primary source of information, where the reader is overloaded with inaccurate contents (as mentioned in Chapter 1.1.2), *(ii)* the greater speed with which fake news have been shown to propagate in online contexts with respect to the real news [73], *(iii)* the mnemonic and cognitive gaps underlying the misinformation effect (Chapter 1.1.1), *(iv)* the reluctance of users to revise their belief system in favour of a more realistic view of the news events (Chapter 1.1.2), *(v)* the lack of a strict control that can filter out only posts with verified information or that do not present dangerous narratives on social platforms (Chapter 1.1.2), *(vi)* the influence exerted by other members of the virtual community (as deduced in Chapter 1.1.1 from [9]), and, most importantly, *(vii)* the innate psychological propensity - based on personality characteristics - that results in a greater tendency to produce and disseminate fake news. The latter factor, in particular, turned out to be of great influence for each category of fake news, both in intentional and unintentional cases.

The above dynamics have, therefore, created non-intermediated contexts where the evaluation of credibility is left, in absence of a robust knowledge, to users' judgment, which is however compromised by the difficulty to understand the veracity of posts about unfamiliar topics. This condition led to what has been defined as the 'post-truth era': all the events can be simultaneously narrated from several perspectives, each of which can be commented in real time by the readers. This fosters the development of virtual environments which are saturated with inaccurate information and toxic narratives: the so-called filter bubbles and echo chambers.

2.1 Informational bubbles

It is by now well-known that, during the evolution of the last decade, the digital industry has shaped a new model, often referred to as the 'data economy' [74, 75]: a system in which services are largely free and, therefore, the main income of platform providers is based on the collection and the exchange of the data derived from users' activities. In particular, the biggest Information Technology companies use algorithms able to derive users' interests, based on their search and browsing history, as well as their interaction with other users or directly with the articles[1]. Once profiled, each consumer becomes the subject of a series of small personalised marketing campaigns, and the results of the future searches on the same platforms will be custom-filtered.

Nevertheless, while advanced Information Retrieval techniques can avoid unnecessary items according to user's preferences and make the service more efficient, on the other side they risk overfitting the observed data about user's interests, ending up over-filtering the information requested during navigation. As a consequence, after prolonged use, the user finds himself isolated within an informational bubble: what is offered to him always adheres to the same cultural and ideological belief system [76].

This condition was first conceptualised in 2011 by Eli Pariser under the name 'filter bubble'. A filter bubble is thus defined as a virtual environment composed of information personalised by the algorithms [77], which, in an attempt to predict what information is relevant based on past click-behaviour and search history [78], lead the user into a state of intellectual isolation [79, 80]. This phenomenon is manifested, for instance, in the great distance between the results produced for different users after feeding a Web search engine with the same word. For example, as mentioned by Pariser, for some the keywords 'British Petroleum' may return articles about the ecological disaster caused by the company in 2010 as a result of an accidental oil spill[2], while for others it may return only investment advices [77, 81]. Another interesting experiment mentioned by Pariser is the search for the word 'Egypt' on the Google platform: in addition to the overlapping of the most obvious results - related to topics such as information about the nation or about travelling - some people saw a lot of articles related to the 2011 Egyptian revolution appear on the first page, while others did not access this information [80].

Therefore, once in this status, the individual not only does not access new information that could enrich him or change his mind, but often also takes an aggressive attitude towards opinions outside his own self-interest bubble [82]. This is how the process of polarization and manipulation of

[1] Google, for instance, reveals the data collected and the reasons in the 'Your Data' webpage: https://safety.google.

[2] A description of the 'Deepwater Horizon incident' is available at the link: https://incidentnews.noaa.gov/incident/8220.

readers takes place: in fact, the 'filter bubble' effect makes users more vulnerable to misinformation and exposes them to disinformation, especially the aforementioned categories of propaganda 1.1.2 and other deceiving contents 1.1.2. In particular, these consequences are amplified within communities known as 'echo chambers', which are even indicated by some research [83] as the primary cause of the polarisation of narratives in online environments. The echo chambers can be defined as internet subcultures that act as a sounding board for these 'ideological frames' [84], or threads in which certain beliefs and ideas are reinforced and evolved through constant repetition in closed virtual environments. Another definition is provided by Sunstein [85]: cliques of users who mutually polarize their opinion in a common and usually radical or even violent ideology, starting from the same biases.

Due to the characteristics with which Web 2.0 has been designed to support the 'data economy' model, the echo chambers are developing more frequently and quickly, especially in absence of an expert intermediation as is the case with microblogging services. As a consequence, these bubbles modify the internal composition of the Web towards what is being recently called 'splinternet' [86] or 'cyber-balkanization' [87, 88]: the internet appears split into sub-groups of like-minded people who isolate themselves within their own community and refuse to be open to new views. Thus, the main characteristics of these groups are *(i)* the lack of external bonds, *(ii)* a strong attractiveness that initially produces exponential increases in the number of members - followed, after reaching a saturation threshold, by more gradual ones -, and *(iii)* the tendency to a negative emotional polarity - having also observed that a greater involvement corresponds to a more negative approach [89]. In particular, regarding the latter point, again Vicario et al. [89] conducted an analysis on several echo chambers on Facebook, differentiated according to the topics discussed in two main categories: communities dealing with conspiracy theories and communities promoting scientific news and research advances. Then, the authors demonstrated that, in each, the level of negativity that can be computationally extracted from the textual part of the posts is higher for users with more than 100 comments. This correlation underlies the speed with which users polarise their ideologies within the echo chamber. Finally, it is also shown that this radicalisation pace is greater in groups discussing conspiracy theories than in those dealing with scientific content. Another study proposed by Quattrociocchi et al. [90] point out how the Facebook ecosystem is gradually fragmenting mostly into highly polarized and non-interacting communities, which, regardless of the specific narrative they focus on, have the same statistical properties and similar ways of interacting with new content. In addition, the authors show how a greater activity by the users within the community is reflected in a greater tendency to interact with similar ideas with respect to the ones recurring in the echo chamber. In this paper, it is then again underlined that it is the functioning of microblogging services like Facebook that favours the

tendency to seek and receive information that reinforces one's beliefs and to reject content that tries to refute them: not only posts related to the preferences are proposed, but the platform also suggests profiles with common interests.

With regard to the psychological causes underlying the phenomenon of echo chambers, two main factors can be mentioned. First of all, users tend to give more credit to content that has been recommended by their peers, regardless of its actual reliability. Therefore, within these communities, the credibility of each information is no longer a value to be personally verified, but becomes a matter of trust in the other members. Thus, the concept changes to what is called 'social credibility'. Secondly, drawing on Zajonc's works [9], the 'frequency heuristic' factor can be appropriately mentioned: readers favour narratives to which they are most frequently exposed, even if they are incorrect.

2.2 Fake News Spreaders

Within the virtual environments just described, different types of actors contribute to the expansion of the dangerous phenomenon of disinformation and misinformation. These diffusing agents can be distinguished into different categories.

In some cases, the following classes of fake news spreaders may overlap.

Journalists and Reporters Professionals involved in disseminating information play an important role in ensuring its accuracy, both offline and online. Sometimes, however, they end up favouring the circulation of fake news in two ways: passively, due to errors or incomplete verification of sources and content of articles; actively, when stories are manipulated to be more marketable, with the second purpose of increasing the number of readers [91](as summarised in Paragraph 1.1.2). As an instance, sensationalized reporting of crime news is common [92]. Other examples of bad journalism are frequent when covering issues about international relations between countries such as Russia and United States [93].

Governments and Regimes Throughout history, the maintenance of State power has occasionally been based on the manipulation of citizens' opinion through the propagation of 'lies' - i.e. distortion or misrepresentation of facts -, persistently perpetrated directly or indirectly by governmental institutions [94]. Well-known examples from the 20th century are the stories fabricated during the Cold War - both by the Western and Eastern Bloc - to keep the people in a constant state of psychosis about the enemy [95], or the so-called 'Cross-Strait' propaganda used by the People's Republic of China against Taiwan and vice versa to convince opposing militias to defect and

rebel against their respective regimes [96]. In addition, as explained in Paragraph 1.1.2, recently it has become increasingly common to use social media to spread false information that can divert public opinion in a foreign state and change its internal political dynamics (as in the aforementioned cases of Russian interference in the U.S. presidential election and during the Brexit referendum). This practice especially exploits the advantages of microblogging platforms, including: *(i)* the possibility to collect data from users and produce targeted campaigns, *(ii)* the speed and ease with which messages can spread, *(iii)* fact checkers' difficulties in contrasting such disinformation contents.

Political Entities and Activist Similar strategies to the one presented in the previous Paragraph, namely aimed at convincing the reader with specific narratives, can also be used by non-governmental political entities, such as parties. Their objective then becomes to damage the image of their opponents and celebrate their own, especially during electoral campaigns. As already illustrated in Paragraph 1.1.2, the intense activity of consultancy companies such as Cambridge Analytica has shown how this kind of microtargeted political marketing on social networks - especially Facebook and Twitter - is widely used worldwide to attract uncertain voters and retain supporters. The same can be said for organisations not strictly linked to the political context, such as the National Rifle Association of America (NRA). The NRA advocates gun rights in U.S. and is responsible for fabricating numerous fake news stories aimed directly at damaging the credibility of newspapers such as The New York Times and The Washington Post and the Democratic Party [97]. In detail, the association, from 2017, began accusing such entities of perpetuating a false and dangerous rhetoric which would induce left-wing voters to engage in acts of rebellion, attempting to document these violent episodes through a series of fake videos on YouTube. In addition, the NRA has repeatedly opposed research about the correlation between gun use and mental disorders by sharing fictitious scientific information [98].

Criminal Organizations and Terrorist Groups The development of information technologies in Web 2.0 and the characteristics of social media have also benefited various international criminal organisations. The most popular case is that of the Islamic State of Iraq and Syria (ISIS). The former militant Islamist group was able to engage in a cyber-war through the streaming of professionally edited shocking videos able to target easily influenced individuals, and through the dissemination of messages of hate and direct threats through their official channels, in order to target Western society. Furthermore, Facebook, Twitter and YouTube platforms have enabled the propagation of the group's radical ideology, with the aim of recruiting

new generations of jihadists [99, 100]. This manipulation mechanism was sometimes carried out through the implementation of chat-bots and even a mobile application [101], which were designed to exploit users' psychological vulnerabilities.

Bots A particularly effective practice to influence online audiences is the use of automated algorithms which learn the same biases penalising human judgement when checking credibility, and consequently share multimedia content that exploits these vulnerabilities [102]. More precisely, if sufficiently sophisticated, such softwares, commonly referred to as 'bots', are able to extract the most persuasive linguistic and visual patterns, and to reproduce texts and images that trigger the desired reactions in the final reader - generally tending towards a negative polarity, hence emotions such as disgust, anger and fear [103, 104]. Thus, the fake accounts regulated by these automated algorithms access the Web through the same channels used by real users and, posing as such, influence and direct online discussion [105]. For a more effective disruption, bots are often coordinated en masse to form networks called 'botnets'.

Bots are used for each of the purposes previously mentioned and are divided into different types, according to the tasks they are able to perform: *(i)* at a low level of sophistication, the only action consists in continuously re-posting some specific content or simply executing the 'like' and/or 'follow' operations; *(ii)* some manage to simulate interaction with real profiles or other bots, producing original comments; *(iii)* in some cases the algorithm indirectly learns the most useful linguistic features and emulates the expression of concepts that are persuasive for the target users, as well as posting fake images that may attract their attention and convince them of the veracity of the post; *(iv)* at the highest level of complexity, it is extremely difficult to distinguish them from a real user, and these fake accounts even end up as real 'influencers' with thousands of human followers[3].

Trolls 'Trolls' are defined within online communities as users who aim to annoy and provoke other users, even aggressively, often simply for their own amusement [108]. Generally, these disruptive actions consist in the publication of contents - textual ones, but even more often images and videos - that are controversial, off-topics and violate the rules of moderation of the conversation in the thread. Sometimes, groups of many trolls combine their efforts to spread waves of false information and distort the normal flow of posts on mainstream platforms such as Facebook and Twitter.

Recent studies [109, 110, 111, 112, 113] about the psychology underlying

[3]Probably, the most successful example is 'Jenna Abrams', a fake account controlled by a Russian-designed bot that even managed to be quoted over a long period by politicians and newspapers of international relevance [106, 107]

such negative online behaviour led to the discovery of specific personality traits and disorders typically presented by these users: trolling is correlated with an anti-social tendency, aggressive behaviour, psychopathy, narcissism and even the Sadistic personality disorder (SPD).

Hidden Paid Posters and Sponsored Trolls This category includes human users who are paid to publish false information on different communities about a topic, create new threads, or target specific demographics to influence consumption habits or social trends [114, 115]. Despite being an innovative strategy in the field of digital marketing, this practice can lead consumers into a state of confusion, divert their belief system and degrade the quality of information circulating online.

Users benefiting from fake news Spreading false information can offer a broad range of benefits, depending on the target. Therefore, within this class are included those individuals who fabricate false content for personal gain, which can consist in an economic advantage, gaining consensus, increasing popularity, etc.

Conspiracy Theorists This refers to that class of users who do not publish fake news for personal purposes or upon request by a third party, but because they truly believe in the veracity of the theories and ideas that are part of a conspiracy narrative. Therefore, they share this information because they are convinced of the importance of the content and so that other people can be persuaded.

2.3 Countering misinformation and disinformation

Misinformation and disinformation are therefore extremely dangerous and far-reaching phenomena able to bring about profound changes in the political, economic and cultural framework of any community, and, in this way, to undermine the foundations of societies all over the world, either intentionally or through narrow mistakes. Although most fake news is extremely easy to disprove, due to the factors explained in Chapter 2, the intermediation by domain experts is by no means sufficient to mitigate the huge flow of false information. Therefore, it seems essential to combine human knowledge tools with a computational approach that can effectively and quickly verify the credibility of online content, after extracting the main patterns of the phenomenon under study.

Among the many possible strategies, this project focuses on analysing the personality characteristics that lead people to believe, share or even produce fake news. Furthermore, it seems interesting to enrich this psychological

information with an analysis of the images that are most often associated with users with a lower capacity of judgement when checking the veracity of online content. In the following Sections, the importance of the study on personality traits and of a multimedia analysis of posts will be introduced separately.

2.3.1 Personality Traits and Emotion Analysis

From the previous Sections, in particular 2, it is evident that underlying the expansion of misinformation and disinformation are exposures that stem from recurring personality traits. This is because, since psychological characteristics regulate behaviour and interaction in the real world, it is reasonable to assume that they are also influential within virtual communities. These psychological traits distort users' understanding, makes them differently sensitive to certain emotions and diversely inclined to spreading not only false information but also toxic narratives. For example, summarising the studies mentioned so far, the correlation between fake news and negative feelings is first of all clear. This link has in fact been investigated by Shu et al. [116]: here, the authors highlighted how fake news aims at triggering emotions such as anger and fear, causing doubts, mistrust and irrational behaviors. In addition, some mental disorders appear to be decisive: in previous research, for example, paranoia, insensitivity and aggressiveness recur. These tendencies, Shu et al. [116] note again, generate what are considered to be the two main factors of vulnerability to fake news: *(i)* näive realism, which corresponds to the erroneous certainty that one's own view of reality is the most accurate, while any other interpretation is regarded as irrational and incorrect [117]; *(ii)* confirmation bias, according to which individuals are more likely to take in information that agrees with their current opinion [118].

It is therefore useful to implement computational techniques that define a user profile from the available data, which in the case of microblogging platforms consists mainly of the textual part of the posts. Subsequently, once the psychological description of several users has been extracted, the aim is to search for correlations between personality patterns and the tendency to spread and/or produce false information.

2.3.2 Multimedia Analysis

Social media contents, however, are characterized by being very often multimedia as they incorporate images and videos. Sometimes, the images themselves, if interpreted in a deliberately misleading way, become a source of misinformation. Many such examples exist: in 2016, the frame of a video showing Barack Obama playing ping pong was edited to prove his involvement in the paedophile ring of the 'Pizzagate' conspiracy (already mentioned in Chapter 1.1.2); in the same year, a photograph taken during an armed

protest in Greece in 2012 was used to accuse anti-Trump protesters of stirring up a violent rebellion[4]. As in the latter case, in the so-called post-truth era the fluid circulation of multi-media content has made images totally independent of their original sources [119]. This makes it very difficult to verify the veracity of the statements made about such photographs. Nevertheless, Sontag [120] and Adatto [121] show how the mere presence of images can give legitimacy to an article in the reader's eyes.

Image can also be a powerful marketing tool: as content flows quickly in microblogs, users' attention may be initially attracted by the visual elements of the posts rather than textual. Therefore, it is reasonable to assume that fake news not only embeds real images with the purpose of proving a thesis disconnected from their original source, but that it may also include images that have been modified and/or built *ad hoc* to maliciously persuade the user. To conclude, it is possible that images embedded in fake news, especially if intentionally fabricated, attempt to exploit some precise users' cognitive vulnerabilities and thus present distinct patterns with respect to real news.

Thus, in this project, we have also investigated the impact of these visual patterns - described by a set of features - on users' tendency to spread false information.

[4]The fake news circulated through the image exchange has been debunked by Snopes: https://www.snopes.com/anti-trump-protesters-destroy-america.

Chapter 3

Background and Related Works

3.1 Fake News Detection

The need to develop state-of-the-art methods to verify the credibility of information in the increasingly chaotic online ecosystem has recently made the disinformation detection in social media a popular area of research. Over the years, a large number of strategies have been tried out, mainly based on the construction of a general model which, considering a series of variables observed on each statistical unit, can generate a prediction generally concerning the truthfulness of the single content. The problem can be formulated mathematically as follows: [Fake News Detection] Given a news article A and a set of attributes $\overrightarrow{c_A}$ that represent it (including text, image, etc), the task of fake news detection is to predict whether the news article A is a fake news element or not, i.e., $\mathcal{F} : \overrightarrow{c_A} \rightarrow \{0, 1\}$ such that,

$$\mathcal{F}(A) = \begin{cases} 1, & \text{if } A \text{ is an element of fake news} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where $\mathcal{F}$ is the aforementioned model, namely the prediction function that must be learned in order to tackle the task.

Therefore, disinformation detection in social media is formulated as a binary classification problem. This is because a false news item is regarded as a distortion of information made by the content creator: in the general literature on media bias [125] these types of distortion are formulated as a binary problem.

Anyway, regardless of the formulation of the problem, the common framework for disinformation detection consists of two main phases: *(i)* the model definition and *(ii)* the features extraction and selection.

The construction of the model consists in the definition of a prediction function $\mathcal{F}$ or $\mathcal{F}'$, which finds the association between the representation offered by the feature vector - $\overrightarrow{c_A}$ or $\overrightarrow{vU}$ - and one of the two target classes: 'fake news' and 'real news', or 'fake news spreader' and 'real news spreader', depending on the selected task.

The architecture through which this classifier is generated can be defined according to two opposing approaches: knowledge-based and data-driven. In the following Paragraphs, examples of the two architectures will be shown.

3.1.1 Knowledge-Based solutions

This approach, which is less popular in the literature, defines a model 'a priori' and not based on a statistical estimation from the available data - as data-driven architecture do. The a priori model is 'knowledge-based' since the veracity of a content is verified using external sources. This methodology is also known as 'fact-checking' and aims at assigning a truth value to a set of statements within a given context [126], which, in this case, corresponds to news articles in social media. Fact-checking is in turn divided into three categories, depending on the type of external source that provides the necessary knowledge to assess the credibility of the information: *(i) expert-oriented*, *(ii) crowdsourcing-oriented*, *(iii) computational-oriented*.

Expert-oriented fact-checking depends on the activity of domain experts and journalists who, relying on their own knowledge and on other available data and literature, offer a judgement about the reliability of each statement presented in the article. Much of the activity of debunking fake news on the Web is coordinated through specialised websites such as PolitiFact[1], Snopes[2], Lead Stories[3], GossipCop[4] for English news and Maldito Bulo[5] for Spanish ones. This approach, however, in addition to being extremely time-consuming and expensive in terms of intellectual effort, has repeatedly shown its ineffectiveness in virtual environments where information circulates freely and in huge quantities (as already mentioned in the Chapter 2).

In the case of the *crowdsourcing-oriented* fact-checking, all the annotations made by many individuals - even non-expert ones - about a given news item are aggregated within a system to form an overall judgement on the truthfulness of that content. This approach is founded on the social concept of 'wisdom of the crowd', according to which the combination of independent solutions proposed by multiple agents performs better than

[1]The official PolitiFact website is at: https://www.politifact.com.
[2]The official Snopes website is at: https://www.snopes.com.
[3]The official LeadStories website is at: https://www.rand.org/lead-stories-factchecker.
[4]The official GossipCop website is at: https://www.gossipcop.com.
[5]The official Maldito Bulo website is at: https://malditobulo.es.

the majority of individual solutions [127], even when facing problem-solving and decision-making tasks like fake news detection. Some examples of platforms implementing such collaborative system are: Fiskkit[6], where users can propose accuracy scores and comment on the various sections in the news article; the Asian service where anyone can report suspicious information items[7]. Nevertheless, this approach also has strong limitations, as it remains time-consuming and too intellectually-demanding.

Lastly, the *computational-oriented* methodology develops strategies that automatically distinguish false and true information, providing systems that no longer depend on human knowledge or users' collaboration, but that are scalable and time-efficient. The most common application of this architecture is better known as automated fact-checking [13, 128, 129] and is lergely based on the use of external resources. Typically, these sources of knowledge resorted to verifying the information credibility are of two types: *(i)* Open Web and *(ii)* structured knowledge graphs.

In the first case, open resources from the Web are taken as references with which the claims in the news item are verified, in terms of consistency. This mechanism of Information Extraction (IE) has, in the past, required frequent and time-consuming human interventions for the manual specification of both extraction rules and relations of interest. Due to the complexity of model definition, these tools depended on linguistic methodologies - such as dependency parsers and Named-Entity Recognisers - which overfitted the domain of interest. For this reason, the application of these systems was limited to small and homogeneous corpora. Over the years, Information Extraction architectures have, therefore, evolved: now, instead of requiring the a priori setting of the relations searched, they propose a scalable structure aimed at discovering all possible relations in the text. In this context, a leading-edge paradigm is the Open Information Extraction [130], which manages to extract a large set of relational tuples from the corpus without requiring human input or intervention, and allowing the user to explore and name these relationships. One tool that implements this paradigm is 'TextRunner' [131], which supports the extraction and query-based exploration in an efficient and highly scalable way by assigning probabilities to the relations and indexing them. The overall aim of the research in this sector is to approach the so-called 'Holy Grail' [132], which has not yet been achieved and would theoretically correspond to a fully automated fact-checking service able to instantly provide the users with a rating about the accuracy of any claim as it appears in real time. State-of-the-art attempts are: 'ClaimBuster' [133] assigns each sentence a priority score indicating how likely it is to contain an important claim that needs to be verified. In this way, fact-checkers

[6]The official Fiskkit website is at: https://fiskkit.com.
[7]The official website is at: https://grants.g0v.tw.

can concentrate on checking the parts of the text that are 'critical' in this respect; WeVerify [134] is a recent groundbreaking system that evaluates the reliability of all the multimedia parts of the content (text, images, video, etc.) by checking the consistency between the description presented in the selected news item and the descriptions obtained on the same multimedia elements from multiple Web searches.

In the second case, the *computational-oriented* methodologies exploit knowledge graphs: collections of interconnected descriptions on entities and concepts [135], which associate data with a context by means of semantic metadata, and thus allow the user to merge and integrate information items. This type of automated fact checking verifies whether the statements in a given news article correspond to factual elements researched within the network topology of knowledge graph (such as DBpedia[8], or the Google Relation Extraction Corpus). Many works have been proposed with this approach. Shi et al. [136] reformulate the automated fact checking task as a problem of predicting the connection between entities in the knowledge graph: given a claim in the form of a *subject-predicate-object* triple, their solution finds the path which defines the relationship among the three elements and then evaluate the veracity by comparing this result to Web resources. Ciampaglia et al.'s formulation [137], instead, consists in finding the shortest path between concepts within a knowledge graph: the reliability of an information is calculated on the basis of the distance among the mentioned concepts. Shiralkar et al., in [138], illustrates the 'Relational Knowledge Linker' algorithm, that exploits the shortest semantically related path within the knowledge graph. Lastly, Pan et al. [139] propose an innovative framework that *(1)* simultaneously exploits three different knowledge graphs on the same corpus to generate the background knowledge, *(2)* implements a Translating Embeddings model for multi-relational data called 'B-TransE' [140] to extract a low-dimensional representation of the entities and the relationships between them, thanks to which *(3)* it classifies the information item as 'true' or 'false'.

In some cases, the knowledge-based approach for fake news detection on social media does not exploit external resources such as those just presented, but rather a set of multiple features describing each news item. In this research context, in contrast to a data-driven architecture, where the classifier is estimated from the available training set, a known strategy is to formulate the model definition as a *Multi-Criteria Decision Making* problem (MCDM) [141]. In a *Multi-Criteria Decision Making* task, the following are presented: a set of candidate solutions - which in our case of study correspond to the news items -, a set of multiple criteria by which these alternatives are rated - interpretable as the credibility characteristics that are taken into account

[8]The official DBpedia website is at: https://wiki.dbpedia.org.

-, and, possibly, distinct weights of importance associated with each criterion. Solving the MCDM problem involves selecting the optimal alternative, i.e. the one that most closely matches the preferences of the decision-maker. Works implementing this system include [142], where Pasi et al. obtain an overall credibility score by aggregating the performance values assessed for each credibility feature, where these performance values indicate the degree to which each criterion is fulfilled for the given news item. The authors test the potential of the model both by giving equal importance to each criterion and by associating different weights. The selected credibility characteristics pertain to four aspects concerning the given content: *(i)* structural features; *(ii)* user-related features; *(iii)* content-related features; *(iv)* temporal features. Lastly, the aggregation is performed using an aggregation operator (AGOP), which in the case of Pasi et al. belongs to the family of *Ordered Weighted Averaging* (OWA) operators [143]. Functions under the class of averaging operators have often been used in the literature, such as those discussed in [144],[145], and [146].

3.1.2 Data-Driven solutions

The data-driven paradigm is configured as a process of mining knowledge compelled by structured and unstructured data [147], rather than by intuition, personal expertise, collaboration between users or use of organized Web resources. The construction of the model, in this case, consists in implementing Machine Learning techniques for learning the patterns among the features which describe the analysed statistical units. Therefore, unlike most of the knowledge-based solutions seen in the previous Chapter 3.1.1, this architecture bases much of its success on the feature extraction and selection. Specifically in the fake news detection task, the feature extraction phase can be interpreted as the attribution to all the interesting elements of the information item (textual part, visual content, semantic level, user's property, etc.) of a formal mathematical representation. The feature selection stage, on the other hand, focuses on finding the best combination of these representations in terms of the result achieved in the classification step - which is generally measured by metrics such as accuracy and F-measure. In past research, many different types of features have been experimented that could describe the credibility of online content in a *machine-readable* way. Following the literature, these characteristics can be divided into several categories. In detail, combining both the popular categorisations offered respectively by Shu et al. [116] and by Buntain & Golbeck [148] - which drew on earlier work by Castillo et al.[149] - results in a subdivision into two main macro-categories: *(i) content-based* features, i.e. related to the content itself, and *(ii)* features related to the *social context*, i.e. describing the internal dynamics of the virtual platform on which the posts are published. The following paragraphs will present these two categories and some works

in which they have been used. Some projects using a mixed approach will also be introduced.

3.1.2.1 Content-Based Features

This type of feature offers a representation of each element of the individual post: the textual part and any visual content. The textual part, depending on the category of informational item, can be organised in various ways: for example, in a news article, it may be appropriate to analyse not only the body text, but also title, subtitle and captions. Given the 'bimediality' of content, the extracted features can be further divided into *linguistic-based* and *visual-based*.

As thoroughly explained in Chapter 1.1, since fake news are often intentionally fabricated to mislead readers and exploit their cognitive vulnerabilities, it is reasonable to base the detection task on the search for precise linguistic patterns, which in most cases are instigating or cause of confusion. With a *linguistic-based* representation, therefore, we seek first of all an association between certain writing styles and the degree of credibility of the information provided. Its extraction may derive from different levels of the text - characters, words, sentences or documents. Usually, these stylistic features are of two types: *(a)* lexical, when deriving the words, characters or phrases that best help discriminate the reliability of news items; *(b)* syntactic, when focusing on the recurrence of some expressions and on the function of terms within the sentence. Some syntactic approaches are: Bag-Of-Words [150] (in which the occurrence of each word is used to train a classifier), N-grams [151] (where sequences of n contiguous words or characters are collected from the corpus), Part-Of-Speech tagging (the computational process by which we label the lexical category of each word to obtain its function within the context it is used). Popular solutions in literature involving the only stylometric analysis of the textual part of contents in social media consisted in the training of Machine Learning algorithms with linguistic features to identify which style fits the category of 'fake news': Afroz et al., in[152], obtained an F-measure of 96.6% by using features such as quantities of syllables and words, vocabulary and grammatical complexity and Part-Of-Speech tags for the classification of documents into 'false' and 'real'; in [13], special patterns in the use of personal pronouns and swear words become an indicator of less credibility. Mendoza et al., in [153], analysing tweets about the 2010 Chilean earthquake, showed that misleading posts are characterised by a higher proportion of negations and contradictory expressions, and at the same time by a less variable vocabulary. Furthermore, in this work, the authors exploit quantitative variables such as the frequency of question marks, exclamation points, first/second/third-person pronouns, and emoticons. Recently, the implementation of more complex architectures for natural language processing has become increasingly popular. For in-

stance, Wang [154], after collecting thousands of labelled news items from the PolitiFact website, exploits Convolutional Neural Networks (CNN) to assess their credibility. In particular, at a syntactic level, models such as Deep syntax [155] and Rhetorical structure [156] can describe syntactic structure by mining deeper rules that better detect the presence of deceptive narratives.

From the studies presented above, it is clear that fake news aims to engage the audience by trying to provoke certain negative emotions, such as disgust, anger and fear. For this reason, the systematic use of polarised language patterns is often considered as a factor of low credibility, because instead of presenting the news in a neutral way, this sensationalist strategy ends up leading readers to misinterpretations. Emotions are mined from the text by means of Sentiment Analysis techniques, which more frequently consist of lexicon-based approaches that associate a prefixed score to each term in the corpus, according to the emotional dimensions it conveys. Some well-known external resources that allow such a method are: the *NRC* Emotion Lexicon [157], describing about 14000 unigrams with eight basic emotional dimensional (*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy*, and *disgust*) and two sentiments (*negative* and *positive*); *SenticNet* [158], providing an emotional tag for more than 200000 linguistic concepts according to a set of vector representations about both semantic and polarity dimension; the lexical resource *SentiWordNet* [159]. In the analysis of the internal dynamics of so-called echo chambers proposed by Vicario et al. [89], and already cited in the Chapter 2.1, the authors manually annotated some example of Facebook posts inside these communities and than trained a Machine Learning algorithm on them to perform an automatic supervised sentiment classification. In doing so, they come to the conclusion that a correlation exists between users' involvement and a negative emotional polarity, whether they dealing with scientific topics or conspiracy theories. As an alternative, Giachanou et al. in [217] operated a document classification into 'fake news' and 'real news' with a deeper architecture: a Long Short-Term Memory network, trained on emotional dimensions defined from the observed lexicon. Then, an emotional insight about false information in social media is offered by Ghanem et al. [160].

In other instances, it is objectivity that is regarded as the main factor underlying the credibility of a news item. Thus, here, the data-driven strategies have the purpose of detecting signals of a a lower degree of objectivity in the text, which could distort the readers' opinion, as in the case of hyper-partisan views or sensationalist press. For example, Nakashole & Mitchell [161] developed an objectivity supervised classifier, trained on a labeled dataset from the Amazon Mechanical Turk study[9].

Lastly, with regard to the class of *linguistic-based* representations, it is worth

[9]Official website of the Amazon Mechanical Turk at: https://www.mturk.com.

mentioning the sub-category of so-called 'domain-specific' features, referred to the use of citations and quotes.

In Chapter 2.3.2, it is well emphasised that visual elements may capture readers' attention and distort their judgement and attitude, sometimes even more than the textual content itself. This is because images and videos are also able to capitalise consumer weaknesses, eventually activating equally powerful emotional responses. At the same time, though, it has been repeatedly shown that images offer fundamental clues for the detection of fake news on the Web 2.0. The visual clues proposed in the literature for tackling this task can be classified into four categories: *(a) forensics features*, *(b) context features* , *(c) statistical features*, *(d) semantic features*.

As regards the first class *(a)*, the goal consists in directly assessing the truthfulness of the news item by verifying that the photographs embedded are original, i.e. that they have not been subjected to modification and manipulation, nor have they been created by implementing deep models (as is the case with so-called 'deep fakes'). These forensic visual features are extrapolated from the following perspectives:

- the manipulation detection, i.e. the search for clues of discontinuity produced by editing operations, such as the insertion of one part of an image into another, or displacements within the same item; this type of research has been proposed, for example, by Boididou et al. [162]; in other cases, specific analyses have been conducted to identify such changes: Mahdian et al., in [163], studied the inconsistency of noise within the bi-dimensional signal; Muhammad et al. [164] exploit representations provided by both Steerable Pyramid Transform (SPT) [165] and Local Binary Pattern (LBP) [166] methods;

- the generation detection, which corresponds to the collection of evidences on the use of deep generative neural networks for the creation of the given image (like the popular Generative Adversarial Network GAN [167]); the visual features derived from this perspective were used for the fake news detection task in the following cases: Nataraj et al. [168] extracted them using a Convolutional Neural Network, McCloskey et al. [169] instead started from the assumption that false images have different correlation patterns between the three *R-G-B* colour channels;

- the re-compression detection: since fake photographs undergo multiple compression due to re-saving after manipulations, re-compression clues are assumed to be effective information for the classification task, as demonstrated in works like [170], [171] and [172].

Contextual features offer quantitative representations for verifying whether the news item with which the given image is presented corresponds to the

original event for which it was taken. In particular, excluding metadata[10] - which are instead used for manual fact-checking - these features are mainly derived from a reverse image search on the Web[11]. From this process we obtain the following: timespan (temporal distance between the publication of the news item and the first publication of the image, which, if higher than a certain threshold, highlights the inconsistency of the visual content); inter-claim similarity (equal to the similarity between the caption of the image in the given news item and the claims crawled from the reverse search engine for the same item).

The *statistical features* derive from the study of the differences between the distribution patterns of images supporting false information and those presented together with reliable news items. Among the most frequently used variables in the literature [173, 174, 175], the following can be mentioned: *(1)* count of images per article or ratio of visual content per news item; *(2)* popularity of the image, which depends on the number of shares or comments it has received on social media, in order to define the so-called 'hot image ratio': the proportion of popular images out of the total visual content posted with the given news item; *(3)* characteristics relating to resolution and style are also considered as statistical features; *(4)* Coherence Score, calculated as the average similarity between any pair within the set of images referring to a given news item; *(5)* Similarity Distribution Histogram, which consists in a vector representation of the similarity relations between all the images within the same set, obtained by mapping the similarity matrix into a histogram; *(6)* Diversity Score, which is the difference between all images within the news set, intended as a complement to the similarity; *(7)* Clustering Score, the number of groups resulting after clustering by visual similarity the images included in the news set.

Finally, in order to exploit the invidual vulnerabilities and spark an emotional involvement, it is assumed that photographs attached to false information possess different semantic characteristics than the images embedded in real news. Convolutional Neural Networks are a very powerful architecture for extracting efficient vector representations of the visual *semantic features (d)*. In this context, some works, instead of training the models on the available data, exploit a transfer learning approach [176]: assuming that, in general, the knowledge extracted from a deep model while solving an original problem can be transferred without significant loss of effectiveness on a new task, a pre-trained neural network is directly used to perform the classification on new data. For example, several applications exploit CNNs that have been pre-trained on the popular ImageNet dataset[12]: this is also done by Giachanou et al. [123], which is the main reference paper of this

[10]When available, metadata of visual content provide relevant information about the file itself, as well as about its production - such as location and time.

[11]An instance of reverse search engine is by Google: https://images.google.com/.

[12]Official Websit of ImageNet dataset at: http://www.image-net.org.

work regarding the stage of visual information extraction. In other cases, in order to enhance the mining of semantic features, more complex architectures based on basic CNNs have been proposed: Qi et al. [177] design a multi-domain neural network able to combine a more abstract semantic representation with a low-level one, under the hypothesis that the emotional provocations produced by the image involve both visual levels.

3.1.2.2 Social Context Features

Beside exploiting information relating to all elements of the individual content, it is often useful to extend the analysis to a broader perspective, taking into account patterns that define user trends within the virtual social contexts. These attributes are thus linked to the forms of interaction allowed by social platforms. Focusing on three main aspects, they can be presented in the following sub-classes: *(a) user-based* features; *(b) post-based* features; *(c) network-based* features. The examples of features illustrated below are originally inspired by the aforementioned work of Castillo et al. [149] and have been reused by the same authors in [178], Buntain & Golbeck in [148], and Tan et al. in [179].

User-based are extracted on two separate levels: individual-wise or group-wise. In the first case, these descriptors detect data from individual users that could be correlated with their tendency to spread fake news on social networks. Some examples are: demographics - such as range of age, location, etc. -, number of followers, number of followees, number of tweets posted, verified 'status' - consisting in a profile's authenticity degree ascertained by the platforms according to its activities -, the density of its virtual social network - measured, for example, as the number of nodes in a graph representation of such interactions -, the time during which the account has remained active since its first tweet. Then, starting from the assumption that fake news spreaders tend to form communities - like echo chambers - and assuming that these communities possess precise characteristics, the credibility of a news item can also be assessed by aggregating the features of the individual users who published it. This is the case with the same group-wise approach adopted by Yang et al. in [180], Kwon et al. in [18] and Ma et al. in [181]. Hence, the same values mentioned above are collected at community level and converted as follows: if the original variable is binary, such as account authenticity, the final representation is derived from the percentage of users with that characteristic; otherwise, averaging or weighting operations are performed.

The so-called *post-based* features, to assess the reliability of a news content, consider how multiple users interact and react towards the given informational item. First of all, the social response can be computationally

studied from a linguistic-based point of view. Ruchansky et al. [182], for example, exploit a word embedding representation to extract semantic features for user comments in the thread concerning the news item. Jin et al. [183], instead, relied on the so-called stance features sub-category: in order to check the truthfulness of an online news article, focus on mining the readers' opinion, classifying it as supporting or denying. For the same task, again Ma et al. [181] extract the topics resulting from monitoring the discussion in the thread.

In addition to linguistic-based methods, *post-based* features can also be derived from a temporal analysis of the thread elicited by a given information content. In detail, the feasibility of predicting the credibility degree of a content is tested by monitoring the temporal variations displayed by the social response with respect to it. A basic feature of this type is the 'lifespan', corresponding to the time elapsed between the first tweet published in the thread and the last one. Other more complex measures have been proposed in the literature: Ma et al. [184] and Ruchansky et al. [182] utilise the Recursive Neural Network architecture to identify variations and anomalies in the discussion over the time. Once a time series has been obtained that can describe the temporal progress of the discussion, some metrics can be obtained for the fake news detection task, like the set of parameters presented by [185]. In particular, it was shown that this temporal approach is particularly effective in identifying rumours in microblogging platforms [186]. Finally, less complex approaches make use of statistical features collected directly from the thread in which users interact about the given news item. Works already mentioned in this Chapter [149, 178, 148, 142, 179] build variables by counting tweets that contain structural elements: multimedia - images and videos -, mentions, URLs, hashtags. Other statistics related to the thread are: number of retweets, number of exclusively textual tweets and average number of tweets posted in the timeline of the users engaged in the discussion.

The last class of features related to the social context is the *network-based* one. In this case, the credibility of a post is estimated by examining the properties of the networks of users who created it, re-shared it, or at least interacted with it. Depending on the social aspect to be investigated, different types of networks are theoretically built, from which in turn derive different representations that are useful for tackling the task of fake news detection. The types of networks popular in the literature are listed below:

- stance networks, where the nodes represent the tweets directly related to the news item, and the edges joining them are associated with a weight equal to the similarity between the expressed opinions (extracted with the same techniques used to define the stance features mentioned above); this type can be found in Jin et al. [183] and Tac-

chini et al. [187];

- co-occurrence network, in which the similarity between users depends on their engagement towards the same information items, i.e. on whether they have both published posts referring to it; Ruchansky et al. tested this methodology in [182];

- friendship networks correspond to the virtual links between the users considered - 'friends' or 'followers', depending on the social platform -, like in [185];

- diffusion network, an evolution of the previous structure as it also incorporates important information concerning the dissemination path[13] of the news within the friendship network.

Once these networks have been defined, the network-based features are derived from the calculation of well-known metrics from Graph Theory research: clustering coefficient and degree coefficient, especially for the categories of friendship and diffusion network.

3.1.2.3 Mixed Approaches

Finally, some studies founded in data-driven architectures have combined some feature types defined under the two separated macro-classes presented in this Chapter (i.e. *(i) content-based* and *(ii) social-based*): Qazvinian et al. [124], besides of content-based features, exploit also network-based features (like, for example, measures about users' tendency to retweet each other's posts), as well as Twitter-specific variables related to hashtags and URLs. Conroy et al. [188], instead, attempted to effectively mix some linguistic-based cues related to individual contents with behavioral data about social dynamics.

3.2 Author's Profiling: Detecting Fake News Spreaders

Considering the mathematical formulation of fake news detection task (Definition 3.1) expressed in the beginning of Chapter 3, it is easy to reformulate the problem from an author's profiling perspective: once the news articles posted by a single user have been labelled as 'fake news' or 'real news' by using any of the data-driven approaches shown in Chapter 3.1.2, the same user can be classified as a 'fake news spreader' if he/she has posted - for

[13] A diffusion path exist between two users a and b only if at least a follows b and a started to publish content about a news item just after b.

example - at least one fake news item[14]; otherwise, the author is considered as a 'real news spreader' up to that point. The mathematical formulation can be expressed as follows: [Fake News Spreaders Detection] Given a user U and a set of attributes $\overrightarrow{v_U}$ which describes certain aspects of him/her (such as personality traits, stylistic characteristics, properties of his/her social context, etc.), the task of fake news spreaders detection is to predict whether the user U presents the tendency to spread elements of fake news or not, i.e., $\mathcal{F}' : \overrightarrow{v_U} \rightarrow \{0, 1\}$ such that,

$$\mathcal{F}'(U) = \begin{cases} 1, & \text{if } U \text{ posted elements of fake news} \\ 0, & \text{if } U \text{ is a real news spreader} \end{cases} \tag{3.2}$$

where $\mathcal{F}^{'}$ is another prediction function used to perform the authors' profiling task.

In general, therefore, for an authors' profiling perspective, Machine Learning models are trained on variable which may simply coincide with the aggregated *user-wise* version of the same features seen when assessing the credibility of an individual information content. The type of aggregation depends on the properties of the given feature. For instance, in the case of content-based linguistic features, it is common to consider the set of texts produced by the user as a single corpus, from which the patterns useful for the classification into 'fake news spreader' or 'real news spreader' are then extracted. Such a strategy was initially proposed for the profiling of blog authors and formal text by Argamon et al. [189] and Koppel et al. [190]; later, due to the very strong diffusion of social networks, similar solutions were also experimented on texts coming from microblogs, thus informal, as done by [191] and [192].

As an alternative to this semantic insight, from a series of n posts published by a single author, it is useful to collect statistical features related to style. For example, Johansson [193] trained a Random Forest model on measures like the tweets length - in terms of word and character -, the number of capital and lower letters, or the number of user mentions.

Starting from the same considerations, , as will be shown in the Chapter 6, to extract a unique representation of visual information at a user level, it is possible to aggregate all the visual feature vectors extracted for each image with operations such as averaging, and then analyse the correlation between this new variable and the tendency to disseminate false information. This solution still appears to be under-explored in the research literature and is one of the main contributions of this work.

[14]Different thresholds can be applied to the number of fake news items published for a user to be defined as a 'fake news spreader'. In this project, as specified, the chosen threshold is 1, so an author is labelled as 'fake news spreader' if he/she has shared at least one post containing information verified as false.

Either way, in this work, the main reference for the task of profiling fake news spreaders is what has been seen from the data-driven systems presented at the PAN 2020 shared task. Hereafter, besides introducing the task, we will describe the creation of the corpus on which the participating systems were tested (Chapter 7) - and on which the experiments related to this project have been carried out. In the final Paragraph, the focus will be on presenting the two best solutions, which it is reasonable to consider as state-of-the-art models.

3.2.1 PAN 2020: Profiling Fake News Spreaders on Twitter

The 2020 edition of the PAN event came with a shared task [240] aimed to inquire the feasibility of detecting authors who shared fake news in their past timeline, especially when they do not only retweet specific news items. The task has been proposed from a bilingual perspective, considering both English and Spanish tweets.

In the next Paragraphs, the following topics will be discussed respectively: the dataset generation and the modifications made in order to carry out the experiments using visual features, the approaches generally used by the participants and the two best-performing solutions.

3.2.1.1 Dataset and Evaluation

The organizers provided two separate databases for each language. Below is the corpus extraction process. Firstly, they selected news certainly labelled as fake on debunking websites[15], i.e. platforms where journalists and experts carefully check the veracity of each statement within the articles, in order to label them as "true" or "false" information. Then they downloaded all the tweets containing information directly related to these labelled fake news items. After removing the tweets not directly referring to the selected topics, they manually labelled the remaining ones as supporting the fake news or vice versa - when authors warn about the untruthfulness of the content. At this point, the manual annotation of the users in the sample has been performed: if one had shared at least one tweet supporting a fake news, then it was labelled as a 'fake news spreader'; otherwise it was annotated as 'real news spreader'. The users were then sorted by the number of fake news re-shared, and only the 250 with the highest ranking were included. The datasets have been balanced by adding 250 randomly selected real news spreaders. Thus, to generate the corpus, the last 100 tweets have been stored from each user's timeline, discarding those directly related to the fake news considered above, in order to avoid that the text classifiers were biased towards them. Finally,

[15]The websites are: PolitiFact, Snopes, Lead Stories, GossipCop for English news and Maldito Bulo for Spanish ones.

the datasets have been splitted into training and test sets following a 60/40 proportion (Table 3-I).

Table 3-I: Respective size of training set and test set provided at the Author Profiling Task at PAN 2020, i.e. number of authors in each set.

Language	Training Set	Test Set	Total
English	300	200	500
Spanish	300	200	500

As already explained, this data-base has been used to perform the experiments presented in Chapter 7 and answer the three research questions. Thus, since this work also aims to assess the effectiveness of visual information for profiling fake news spreaders, the corpus was integrated with the patterns identified within the images embedded in the tweets. The retrieval of those images and the extraction of the visual features are described in Chapter 6.

The averall performance of each of the 66 participating systems has been ranked by the average accuracy per language, i.e. the average number of correct answers - correct classifications in 'fake news spreader' or 'real news spreader' - for both the English and Spanish corpora, out of the total number of observations presented in both cases.

3.2.1.2 Overview of Submitted Approaches

Considering average measures between the two corpora (English and Spanish), the mean and median accuracy achieved by the 66 participants are 0.7039 and 0.7125 respectively, with a fairly low coefficient of variation (around 0.07).

As the data set provided by the organisers contained the raw text of the tweets directly downloaded via the Twitter API[16], a pre-processing phase was essential before feature extraction and model training. The most frequently used methods included the removal of various elements from the text: numbers [206, 222, 194, 236, 195, 196], non-alphanumeric characters - including emojis and emoticons - [205, 206, 222, 194, 236, 232, 226, 196], punctuation signs [222, 223, 194, 236, 195, 226, 196]. Other techniques include: lower-casing [205, 206, 222, 225], stop word removal [222, 223, 194, 236, 195, 226, 196], stemming or lemmatization [194, 195, 226, 196], tokenization [222, 197, 224, 236, 226, 196, 234], removal of short text or infrequent terms [222, 198].

[16]Official Website of Twitter API: https://developer.twitter.com/en/docs/twitter-api.

The classification approaches are all, obviously, data-driven. In most cases, the architectures used are based on classical statistical or Machine Learning models: Support Vector Machines [206, 222, 223, 236, 224, 195, 226, 201, 200] is the most prevalent, followed by Logistic Regression [205, 222, 223, 199, 225, 201, 227], and Random Forest [230, 236, 195, 201, 202, 227]. Ensemble methods, in addition to the best solution on the English corpus (Buda Bolonyai [205]), have been built by stacking the same classical models mentioned above with others - such as XGBoost, Decision Trees, Naive Bayes, etc. -, as done in [198, 196, 228]. Less frequent submissions suggested deeper model, including Multilayer Perceptron [201], neural networks with fully-connected layers [194], Convolutional Neural Networks [231], Long Short-Term Memory [232] - also implementing the self-attention mechanism [203] -, and Gated Recurrent Unit (GRU) [204].

Given the setting of the Author Profiling task, the features used are linguistic-based and lexical-based, extracted directly from the sub-corpuses formed by the user-wise union of tweets. Both in the English and Spanish case, all participating classifiers have been trained on features that can be divided into four categories: *(i)* n-grams at a character or word level; *(ii)* stylistics; *(iii)* embeddings; *(iv)* personality and emotions. The most popular solution appears to be that based on *(i)* n-grams: Espinosa et al. [236] employ this representation on both levels - terms and characters. In particular, a common approach is to apply weights to the n-grams, using the Term Frequency-Inverse Document Frequency function (TF-IDF): this is done, for example, by Vogel et al. [222], Koloski et al. [223], Fernández et al. [224] and Pinnaparaju et al. [225]. With regard to the *(ii)* stylistic variables: Manna et al. [227] collects statistics about the occurence of structural elements of posts and complements with an analysis of logical parts of sentences expressing personal opinions; following this second method, Niven et al. [228] monitor the use of adverbs, impersonal and personal pronouns, and function words; Russo et al. [229], instead, in combination with statistics on tweet elements, extracts a measure of lexical variability through the Type Token Ratio (TTP); again, Cardaioli et al. [230] employ ten features describing the writing style. *(iii)* Embedding representation has always been used in combination with other feature categories, as done by Hörtenhuemer et al. [199], Justin et al. [200], Majumder et al. [232], Wu et al. [235]. Lastly, as far as *(iv)* personality and emotions, an interesting work is the one presented by Espinosa et al. [236], who extracted psychographic information through the Symanto API[17].

However, in general, most submitted systems exploit a mixture of these feature classes. In particular, the works that obtained the best result re-

[17]Official Website of Symanto API at: https://www.symanto.com/api/?redirect-from=developer.

spectively exploited a combination of n-grams and stylistic features (Buda & Bolonyai [205]) and only n-grams (Pizarro [206]). These solutions will be better explained in the next Chapters 3.2.1.3.

3.2.1.3 Best Performances

The best average performance has been achieved, with equal merit, by Buda & Bolonyai [205] and Pizarro [206]. In Table 3-II, we reports that Buda-Bolonyai's system provided the highest accuracy on the English corpus, while Pizarro obtained the best result on Spanish tweets[18].

Table 3-II: Best performances recorded in Author Profiling Task at PAN 2020.

System	English	Spanish	Average
Buda-Bolonyai	**0.750**	0.805	0.7775
Pizarro	0.735	**0.820**	0.7775

Such systems can be considered as state-of-the-art classifiers for the task of fake news spreaders detection. For this reason, since the third objective of this paper is to demonstrate that information on personality traits and visual features can improve the performance of best-performing models, experiments were conducted modifying these architectures proposed by Buda & Bolonyai [205] and Pizarro [206].

In the following Chapters the two systems will be separately depicted.

[18]Complete list of the results at https://pan.webis.de/clef20/pan20-web/author-profiling.htmlresults.

Chapter 4

State-of-the-art models

Both systems proposed by Buda & Bolonyai and Pizarro can be considered as state-of-the-art models for tackling the task of fake news spreaders detection. Since the objective of this project is to evaluate the potential of personality and visual features to improve the performances of the best classifiers in this research field, it is appropriate to describe these architectures thoroughly, in order to later understand the variations made during the experiments (Chapter 7).

In addition, it is worth outlining the CheckerOrSpreader architecture, which has been used to perform part of the tests and that inspired the methodologies of personality features extraction proposed in this paper.

4.1 Buda & Bolonyai's System

Buda & Bolonyai's solution [205] can be summarized by the following architecture: five sub-models are separately trained to determine the probability, for each user, of being a fake news spreader, and then they are stacked together through an ensemble method consisting in a Logistic Regression. The five baseline classifiers are: *(i)* a regularized Logistic Regression, *(ii)* a Support Vector Machine, *(iii)* a Random Forest, *(iv)* a gradient boosting algorithm run via the XGBoost, *(v)* another XGBoost model, trained on a different features set. As just mentioned, the predictions are based on a combination of two types of textual variables: sub-models from *(i)* to *(iv)* are trained on n-grams, while the second XGBoost classifier considers seventeen descriptive statistics. The next sub-sections introduce separately the two types of baseline used: the *n-grams models* and the *user-wise statistical model*.

4.1.1 *N-grams* Models

As for the *n-grams* classifiers, the training phase has been set up as the search of the optimal combination among text pre-processing methods, vectorization techniques and the parameters of Machine Learning models, through an extensive grid research.

Regarding the pre-processing stage, two text preparation methods were tested by the authors:

- ***M1***: removing non alphanumeric characters and put to lower case;
- ***M2***: removing non alphanumeric characters except emoticons and emojis and put to lower case.

With regards to the corpus vectorization, different ranges for n-grams on words have been experimented - unigrams, bigrams, mix of unigrams and bigrams - and also the overall minimum frequency of n-gram in the documents has also been optimized, considering starting values in the grid between 3 and 10[1]. Once collected and possibly filtered by frequency, the n-grams can be interpreted as the dimensions on which the Machine Learning models are trained. Anyway, the values assigned to the users at each n-gram are not derived from the raw occurrence within the document - i.e. the number of times these n-grams are found within the corpus formed by the union of the last 100 tweets -, nor the relative frequency. This choice stems from the fact that this measure is biased towards high frequencies and thus ends up associating high relevance values even to terms that are present throughout the collection. Instead, the Term Frequency - Inverse Document Frequency (TF-IDF) function is used. Defined as $n_{i,j}$ the occurrence of an n-gram i within the corpus of the j-th author d_j, the TF-IDF weight is calculated with the equation below:

$$\text{tf-idf}_{i,j} = \frac{n_{i,j}}{|d_j|} \cdot log_{10}(\frac{|D|}{\{d : i \in d\}}) \tag{4.1}$$

where: $|d_j|$ is the length of the j-th corpus in terms of words; $|D|$ is the total number of corpora - or users; $\{d : i \in d\}$ is the number of documents in the collection containing the n-gram i. Therefore, the final TF-IDF value derives from the product between the relative importance of the n-gram within the single corpus and the global relevance in the collection (which can be interpreted as a measure of specificity of i, inversely proportional to its popularity).

Next, it is then worth analysing in depth the characteristics of the Machine Learning models implemented.

[1] In details, the values considered in the search grid for the overall minimum frequency of n-gram in the documents are in the interval (3,4,5,6,7,8,9,10).

Logistic Regression. The Logistic Regression (LogReg) [207] defines a relationship between a categorical target variable Y and a set of prediction features, by evaluating the probability that each independent variable fall into certain levels of the categorical response. In the research branch of Natural Language Processing, the problem is hence formulated as follows: given a set of classes for the target variable $Y = \{y_i | i \in [1, N]\}$ and a collection of documents $X = \{d_j | j \in [1, N']\}$, logistic regression calculates the probability of a specific target class y_i to include the given a document d_j: $p(y = y_i | d = d_j)$. In the specific case of the author profiling task proposed in this paper, the model will calculate the probability of the "fake news spreader" class - complementary to the probability of the "real news spreader" class - given a user, based on the set of his/her characteristics (linguistic-based or otherwise). This probability is calculated with a sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$. During the training phase, this function maps each observation (user) of the training set to a class of the target variable; in this way, the model learns the vector of weights w to be applied to each feature and the bias e, based on which the regression plan is constructed. These weights can be interpreted as the degree of relevance of each independent variable. For instance, in the simple case of a lexical approach, the model will extract the lexical patterns typical of a certain user category by learning the importance of each term from the corpus. Otherwise, given a vector encoding the individual's personality traits, the model will extract the relative influence that these psychological dimensions have on the tendency to disseminate false information or not. So, in conclusion, Logistic Regression is defined by the following equation: given a user u_j defined by a features set x_j,

$$p(y = y_i | x = x_j) = \frac{1}{1 + e^{-(wx_i + b)}} \tag{4.2}$$

The optimal vector of weights w^* is estimated by maximising the log likelihood using the gradient descent algorithm [208].

As will be verified by the results of the experiments (Chapter 7), the Logistic Regression proves to be a more accurate model than others in presence of a correlation among the independent variables, as is the case with textual features and vector representations encoding visual and personality information. In particular, in order to better handle the strong multi-collinearity among features, during the training phase the *shrinkage* statistical technique is applied, which allows to contract the estimates of the regression coefficients towards zero by means of a penalty. In this case, this regularisation consists in the Euclidean norm L_2 [209], which is based on the regularization parameter (C). As said, the value of C is optimised through a grid search, within a range specified in Table 4-I.

Support Vector Machine. The Support Vector Machine (SVM) [210] is a robust supervised method providing a non-probabilistic binary linear clas-

sifier, able - in our specific case - to map the observation related to the users into a new space, according to the representation obtained with the features, and classify them into "fake news spreaders" or "real news spreaders" by maximising the width of the linear distance between the these two target classes. The statistical units from the dataset are mapped into the hyper-dimensional space by kernel functions. Buda & Bolonyai, in particular, use a linear kernel function to conduct the experiments on both the Spanish and English corpora. Hence, the resulting hyper-space will be a linear projection of the points associated with each user, whose original position depends on the TF-IDF weights calculated for every n-grams.

As in the case of Logistic Regression, in order to properly handle the correlation among the textual features - i.e. TF-IDF weights associated with the n-grams -, a regularisation with Euclidean norm is applied. The regularisation parameter C for the Support Vector Machine is included in the set of hyperparameters undergoing the grid search optimisation process, this time in a different interval (Table 4-I).

Random Forest. The Random Decision Forest (RF) [211] is an ensemble method which builds multiple Decision Trees from the labelled data in the training set, and returns a classification prediction corresponding to the class that coincides with the mode of the results obtained from each individual Decision Tree. The core of this model is the application of a *bootstrap aggregating* technique - also called *bagging* - during training stage, which makes the prediction more accurate, reduces its variance and lowers the risk of overfitting [212]. The bagging mechanism, thus, consists of combining several model which are inherently characterised by statistical noise but that are also unbiased. Decision trees are well suited to the bagging technique, because, despite being noisy, they are able to incorporate complex relational structures from the data with relatively low bias. That said, at a basic level the construction of the ensemble architecture is carried out through the operations below: given N observations - or users -, given M independent variables - which, in our case, correspond to all the n-grams identified in the texts collection -, each node of the global tree is trained with m variables randomly extracted with replacement from the original features set on a portion of the dataset with dimension n, and its error is calculated on the remaining $N-n$ observations. In detail, the partition of size n is selected by maximising an index that computes, according to the m randomly chosen features, the quality of each possible sample. During the inference phase, the label assigned will be that of the class most voted by the subtrees.

In the case of this model, there are two hyper-parameters to be optimised (Table 4-I). The first one is the number of trees into which the ensemble method Random Forest is divided. The second one consists in the minimum number of cases on each subtree. Consequently, only split points generating

two partitions with a minimum threshold of users are considered. Finally, the quality of the splits is calculated using the Gini impurity criterion: an index assessing how often a randomly selected observation from the dataset would be misclassified following the distribution of the labels in the evaluated partition.

XGBoost. The term *XGBoost* (XGB) [213] refers to a software for implementing a gradient boosting method in a variety of programming environments, including Python[2] (selected in the context of this work). The Gradient boosting consists in a Machine Learning technique which combines individual weak and noisy predictive models, such as Decision Trees. For this reason, this algorithm can be considered as a more powerful evolution of the Random Forest [214]. More precisely, they are defined as methods which adapt the loss function to optimise the result, by iteratively choosing a prediction function that follows a direction with a negative gradient - called 'weak hypothesis' - from a space of functions. Thus, in a supervised classification task in which we consider a categorical variable y jointly probability-distributed with a set of features X, the gradient boosting strategy returns the approximation of the prediction function that minimizes, in the training set, the expected value of a specified cost function. This approximation is computed as the weighted summation of the M weak hypotheses $h(X)$ - extracted from the same class of functions $\mathcal{H}$ -, i.e. of the individual forecast functions:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(X) + \mathrm{k} \tag{4.3}$$

where k is a constant. The adaptation of the loss function on which the generation of the m-th model at the m-th iteration depends is done by starting with a first constant function of X and performing the expansion visible below:

$$F_m(X) = F_{m-1}(X) + \underset{h_m \in \mathcal{H}}{\arg\min} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(X_i) + h_m(X_i)) \right] \tag{4.4}$$

However, since this greedy formulation of the search of the optimal model parameters is computationally infeasible, the iterative operation of loss function minimization is performed by applying the heuristic of the 'steepest descent' step: at each iteration the prediction function most similar to the gradient of the cost function is chosen for which the step size γ can be calculated. Therefore, following this heuristic, the final model after M iterations will be derived from the following equation:

$$F_m(X) = F_{m-1}(X) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L(y_i, F_{m-1}(X_i)). \tag{4.5}$$

[2]Official library documentation at: https://xgboost.readthedocs.io/en/latest/python/.

When training the XGBoost as one of the baselines within the proposed solution at the Author Profiling task, Buda & Bolonyai select the class of Decision Trees as the function space $\mathcal{H}$ from which the weak hypotheses are extracted (as explained earlier). More over, they perform an optimization of the following hyperparameters of the algorithm:

- *learning rate* η: in order to make the execution of gradient boosting more conservative and thus avoid overfitting, the size of the γ step is shrinked by an η parameter in the range $[0, 1]$, which is recalculated at each step according to the feature set; therefore η can be regarded as equivalent to the regularization parameter C associated with the previous baselines;

- *number of estimators*, i. e. of the baselines - weak prediction functions - extracted;

- *maximum depth of each subtree*, value directly proportional to the complexity of the model, hence the risk of overfitting, as well as the computational and time cost;

- *subsample ratio of the training instances* indicates the percentage of the training set that the algorithm randomly samples once at each iteration, before defining the Decision Tree; this operation is performed to reduce the risk of overfitting;

- *subsample ratio by column* corresponds to the portion of the original features set iteratively randomly sampled at each step when building the subtree.

To make the method even more conservative, a further regularization is applied to the model weights, this time through the L_1 norm. Since, unlike the Euclidean norm, this penalty function is not derivable, the method will tend more to contract the weights to the null value, indirectly performing a feature selection. For this reason, the XGBoost algorithm is able to effectively handle correlated variables, and is therefore perfect for linguistic, personality or visual features (which are vector representations whose values are extracted simultaneously and are often highly dependent). The intensity of this penalty is governed by the parameter α. Lastly, thanks to the advantages offered by the XGBoost algorithm, it is not necessary to normalize the variables.

It is worth mentioning that the high computational cost of this hyperparameters optimisation strategy within the XGBoost model suggested training the algorithm not with all the possible combinations of features but only with the combinations performing the best in terms of accuracy when tested with a lighter model such as Logistic Regression. Otherwise, since the complexity is proportional to the number of input columns, the completion of

the work would have been unfeasible, especially considering the large size of the visual information representation[3].

Table 4-I: Grid-searched hyperparameters for the Machine learning models used as *n-grams* baselines within Buda-Bolonyai's solution at the Authors' Profiling task at PAN 2020.

Model	Hyperparameter	Interval
LogReg	Regularization (C)	{0.1,1,10,100,1000}
SVM	Regularization (C)	{1,10,100,1000}
RF	Number of subtrees	{100,300,400}
	Number of cases	{5,6,7,8,9,10}
XGB	Learning rate (η)	{0.01,0.1,0.3}
	N. of estimators	{200,300}
	Max. depth of trees	{3,4,5,6}
	Subsample ratio	{0.6,0.7,0.8}
	Subsample ratio of columns	{0.5,0.6,0.7}

After performing the optimization, Buda & Bolonyai have obtained the optimal hyperparameters visible in the Table 4-II (for English corpus) and 4-III (for Spanish corpus).

4.1.2 *User-wise Statistical* Model

Finally, the fifth model is estimated using another XGBoost algorithm on 17 user-wise statistical stylistic indicators which can be classified into the following groups:

- ◇ the minimum, maximum, mean, standard deviation and range of the length of the user's tweets in terms of words;
- ◇ the minimum, maximum, mean, standard deviation and range of the the number of characters in the sub-corpora;
- ◇ number of virtual interactions performed by the selected author: retweets and mentions;
- ◇ counting of various structural elements in the text related to the post: URLs, hashtags, emojis and ellipses;

[3]Only taking into account the vectors of visual information, the total number of columns would amount to 7228.

Table 4-II: Optimal combination among pre-processing pipeline (*M1* or *M2*), vectorization technique (choice of the n-gram order and of the minimum frequency threshold) and hyperparameters for each *n-grams* model considered as a baseline within the ensemble solution proposed by Buda & Bolonyai on the English corpus.

Model	Pre-processing	N-gram	Min. Freq.	Hyperparameters
LogReg	*M1*	uni/bi-grams	6	$C = 1000$
SVM	*M1*	uni/bi-grams	5	$C = 100$
RF	*M2*	uni/bi-grams	9	Num. of subtrees = 300
				Num. of cases = 9
XGB	*M1*	uni/bi-grams	8	$\eta = 0.01$
				Estimators = 300
				Max. Depth = 6
				Subsample = 0.8
				Subsample (col.) = 0.6

- ◇ a lexical diversity indicator calculated as the type-token ratio (TTR) of lemmas, corresponding to the total number of unique lemmas - derived from the words used in the given textual segment - divided by the total number of tokens; obviously, the greater this measure, the greater the lexical richness and variability of the user's vocabulary.

As already said, also in this features subset, it is possible to avoid the normalization of the columns and verify the presence of an hypothetical multicollinearity.

The hyper-parameters and parameters of this second XGBoost model resulting from the optimisation originally performed by Buda & Bolonyai are shown in the Table 4-IV.

4.1.3 Ensemble Model

Authors avoided splitting the corpus into a training and validation set, preferring cross-validation techniques to prevent overfitting while optimizing the hyperparameters of the five baselines. To lower the overfitting risk also in the estimation of the ensemble model, its training was not directly performed on the predictions of the five sub-classifiers, but on a new dataset containing an approximation of the predictions distribution obtained by refitting the submodels five times - with the selected hyperparameter- on different chunks of the training set (each one including sub-corpora from 240 users).

Finally, in order to find the best ensemble method, four different techniques were tested on this new dataset: *(i)* Majority Voting; *(ii)* Linear

Table 4-III: Optimal combination among pre-processing pipeline (*M1* or *M2*), vectorization technique (choice of the n-gram order and of the minimum frequency threshold) and hyperparameters for each *n-grams* model considered as a baseline within the ensemble solution proposed by Buda & Bolonyai on the Spanish corpus.

Model	Pre-processing	N-gram	Min. Freq.	Hyperparameters
LogReg	*M1*	bi-grams	9	$C = 100$
SVM	*M1*	bi-grams	8	$C = 10$
RF	*M1*	uni/bi-grams	3	Num. of subtrees = 100
				Num. of cases = 8
XGB	*M1*	uni/bi-grams	8	$\eta = 0.3$
				Estimators = 200
				Max. Depth = 6
				Subsample = 0.6
				Subsample (col.) = 0.7

Regression; *(iii)* Logistic Regression, which has been selected as it turned out to be the most reliable.

The best and most reliable results have been obtained using the Logistic Regression model as ensemble prediction function. Table 4-V shows, for both languages, the coefficients associated with each of the five baselines. As these weights can be interpreted as the contribution of each sub-model to the accuracy of the final classification (as deducible from the Equation 4.2), it is possible to conclude the following from the experiments performed by Buda Bolonyai. First of all, the coefficient assigned to the Random Forest is zero, so such method trained on uni-grams and bi-grams appears completely useless for the task of fake news spreaders detection. Then, with regards to English corpus, the XGBoost trained on uni/bi-grams seems to contribute most, followed by the Logistic Regression and, with a much lower coefficient, the Support Vector Machine. As for the Spanish sample, instead, Logistic Regression and Support Vector Machine clearly outperform XGBoost trained on n-grams. In both cases, the application of a gradient boosting algorithm on stylistic statistics leads to a minor contribution.

4.2 Pizarro's System

The system proposed by Pizarro [206] is configured as a process of optimization of the parameters and hyperparameters of a Support Vector Machine trained on several combinations of word and character n-grams, while experimenting twelve different pre-processing pipelines.

Table 4-IV: Optimal hyperparameters (in the upper part of the Table) and parameters (in the lower part of the Table) for the XGBoost baseline trained on the 17 stylistical features, within Buda-Bolonyai's solution at the Authors' Profiling task at PAN 2020. The results are presented respectively for English (column 'ENG') and Spanish (column 'ESP') corpora.

Parameter	ENG	ESP
Learning rate (η)	0.2	0.3
N. of estimators	200	100
Max. depth of trees	2	3
Subsample ratio	0.8	0.8
Subsample ratio (col.)	0.9	0.8
Descent step (γ)	2	4
Regularization (α)	0.1	0.3

Table 4-V: Weights assigned to each baseline during the training of a Logistic Regression ensemble method originally carried out by Buda & Bolonyai, for both languages.

Baseline	ENG	ESP
LogReg	0.8	**1.31**
SVM	0.48	1.16
RF	0	0
XGB	**1.07**	0.54
XGB (Stat.)	0.2	0.12

As far as the Machine Learning technique is concerned, the model implemented is more precisely a Linear Support Vector Classifier (LinearSVC). This algorithm differs from the original Support Vector Classifier with linear kernel - chosen as one of the baselines by Buda Bolonyai - for using the *LIBLINEAR* library instead of *LIBSVM*. Here are the differences between the two approaches. Firstly, *LIBSVM* [215] applies a Sequential Minimal Optimisation (SMO) strategy in order to solve the Quadratic Programming problem (QP) that typifies Support Vector Machines during the training for a binary classification task. To do so, the binary formulation is broken down into a series of minor sub-problems that can be analytically solved (since, in each of them, it is sufficient to find the minimum of a one-dimensional quadratic function). As a result of this functioning, the complexity of the training of a Support Vector Classifier based on the Sequential Minimal Opti-

misation may be estimated at around $O(n^2)$ or $O(n^3)$, where n is the number of observation - users - in the training set. In contrast, the complexity resulting from the implementation of the *LIBLINEAR* [216] library is $O(n)$, so, as a first advantage, this approach shows a greater ability to scale for large datasets. This different capability arises from the use of a different algorithm when training the model: the Coordinate Descent. This solution consists in the iterative minimization of the loss function with respect to a block of coordinate directions that are selected and subsequently adapted at each iteration according to the step taken along a coordinate hyperplane. This not only reduces the complexity but also gives greater flexibility in the choice of regularization and loss functions. Concerning the first selection, a penalty is applied through the Euclidean norm L_2 to avoid overfitting. On the other hand, the aforementioned flexibility of the approach allows the category of loss function to be included in the list of hyperparameters undergoing the optimization process. In detail, the selection is made between the next two functions:

- the standard Support Vector Machine loss function (*hinge*), defined by the equation:
$$\ell(y) = \max(0, 1 - c \cdot y) \tag{4.6}$$
where c is the binary classification output ("fake news spreader" or "real news spreader") and y is the classification score produced by the model, calculated as a linear combination of the input feature matrix X and the hyperplane parameters w and b: $y = \mathbf{w} \cdot X + \mathbf{b}$;

- the square of the previous hinge loss:
$$\ell(y) = \max(0, 1 - c \cdot y)^2. \tag{4.7}$$

The other optimised Support Vector Machine Linear hyper-parameters are:

- the regularization parameter C
- tolerance value for stopping criteria
- a constant added to the input matrix to scale the intercept of the model
- a binary indicator that performs a balancing of target classes during training by applying weights.

The maximum number of iterations is set at 2000.

Then, different cut-off to the less and more frequent words are tried out as well as different n-grams ranges - searching within a grid between 1 and 3 words and 1 and 6 characters (as visible in Table 4-VI).

Table 4-VI: Combinations of n-gram order and occurrence limits tested during the grid search optimization process performed by Pizarro.

Token	Parameter
Word	N-gram range : $\{(1,2),(1,3),(2,3)\}$
	Max. Frequency : $\{0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$
	Min. Frequency : $[10^{-4}; 5]$
Character	N-gram range : $\{(1,3),(1,5),(2,5),(3,5),(1,6),(2,6)\}$
	Max. Frequency : $\{0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$
	Min. Frequency : $[10^{-4}; 5]$

The final linguistic features derive from the calculation of the Term Frequency - Inverse Document Frequency (TF-IDF) weight for each n-gram (Equation 4.1). This is done by treating the 100 tweets collected for the 1000 users (500 per language) as if they consist in a single document (i.e. a sub-corpus), so as to obtain a unique vector representation.

At the same time, the authors search for the optimal mixture between the textual pre-processing operations included in the following initial set: *(i)* downcase all the letters (*downcase* in the Table 4-VII); *(ii)* replace numbers (*replace_numbers*); *(iii)* replace anonymized tags referred to URLs, users' name and hashtags[4] (*replace_tags*); *(iv)* replace emojis with word representation (*demojify*); *(v)* reduce number of repeated pleonastic characters in a word *(reduce_length)*. The twelve text preparation pipelines tested during the optimization process performed by Pizarro are listed in Table 4-VII.

Thus, in summary, the best solution - in terms of accuracy - proposed on the Spanish corpus at the Authors' Profiling task is based on the following formulation of a problem: the initialization of an optimization hyperspace[5] within which a recursive search is carried out, by trial and error, for the best possible combination of *(a)* a set of text pre-processing operations, *(b)* several vectorization strategies of the textual content of the tweets and *(c)* the parameters of a Linear Support Vector Machine. To improve model estimation, this greedy grid search is combined with a cross validation technique, firstly in 5 folds and then, for a better result, in 10 folds. This 10-fold cross-validation strategy originally led to the selection of the following systems.

For English tweets, the selected text preparation pipeline for both word and character n-grams was ***v2_1_1***: it includes, among the aforementioned

[4] In case this pre-processing operation is performed, the different tokens associated to URLs, users' name and hashtags are respectively: *xxurl*, *xxusr*, *xxhst*.

[5] This is done with *hyperopt* library in Python, whose official documentation is available at: https://github.com/hyperopt/hyperopt.

Table 4-VII: Pre-processing pipelines tried out within the optimization process proposed by Pizarro. The entry of the Table indicates True ('T') if the given operation is included in the pipeline ('Strategy'); otherwise, False ('F').

Strategy	*downcase*	*replace_length*	*replace_numbers*	*demojify*	*reduce_tags*
v0_0	T	T	F	F	F
v0_1	F	T	F	F	F
v1_0	T	T	F	T	F
v1_1	F	T	F	T	F
v2_0	T	T	F	F	T
v2_0_1	T	T	T	F	T
v2_1	F	T	T	F	T
v2_1_1	F	T	T	T	T
v3_0	T	T	F	T	T
v3_0_1	T	T	T	T	T
v3_1	F	T	F	T	T
v3_1_1	F	T	T	T	T

operations, *(ii)*, *(iii)* and *(iv)* with the stop words removal. The linguistic features were extracted from the calculation of TF-IDF weights on character n-grams of orders between 1 and 6, as well as unigrams and bigrams for word n-grams.

As for the Spanish corpus, the best pre-processing strategy turns out to be ***v3_1***: almost the same used for English text, excluding digits replacement. In this case, the TF-IDF index calculation was performed on characters n-grams in a range between 2 and 6, and unigrams, bigrams, and trigrams for words n-grams.

4.3 CheckerOrSpreader

To the aim of mining users' personality traits from posts, this project is inspired by Giachanou et al.'s work [122]. The authors originally proposed a method based on the use of two different vector representations describing psycho-linguistic characteristics inferred from the text, with the goal of classifying English users into 'fake news spreaders' and 'fact checkers' (i.e. those interested to share posts that refute fake news with evidences). Hence, their task differs from the one proposed here: indeed, we aim at a differentiation into 'fake news speaders' and 'non spreaders'.

In order to pursue the aforementioned objective, Giachanou et al. created a database of 2357 users labelled with the two target classes. The construc-

tion of this dataset follows the same process implemented to generate the two samples provided at the Author Profiling task at PAN 2020 (Chapter 3.2.1.1). Briefly, after crawling articles categorised as 'fake' by experts on the LeadStories platform, the headlines of these items - from which the stopwords had been removed - were exploited to perform a search through the Twitter API and to store a sample of tweets relevant to the given news item. These posts were then categorised as 'spreading tweets' - when the text supported the veracity of verifiably false information - or 'fact check tweets' - if it attempted to refute the original article, demonstrating its lack of credibility or mentioning other debunking sites. This classification was performed through a semi-automated procedure. First, rules were defined to manually summarise the linguistic patterns[6] identified within the class of 'fact check tweets', and if such lexical patterns were not found in a text, then the post was considered as 'spreading tweet'. Finally, the users collected from this sample were annotated according to the number of 'spreading' or 'fact check' tweets they had posted: the category with which a given individual is associated corresponds to the modal class in the set of his/her tweets previously labelled. Finally, as in the case of the PAN 2020 dataset, the final corpus - now, in English only - was generated by collecting the last 1000 unannotated posts published by each labelled user.

On this data set, Giachanou et al. then tested various solutions that could predict the users' category according to the personality traits inferred from their writing style. The best performance, an accuracy of 59%, was achieved with the CheckerOrSpreader model. CheckerOrSpreader architecture consists in a Deep Learning-based methodology, whose functioning can be summarised into the following steps. First, an embedding layer is initialised with a semantic vector representation of the text, produced by the Deep Learning architecture called Global Vectors for Word Representation (GloVe) [218]. GloVe extracts global statistical information from the corpus by executing the training phase only on the non-zero elements of a word-word co-occurrence matrix, rather than on the whole sparse matrix (as done, for instance, by approaches such as Singular Value Decomposition) or within a single small window (an operation on which the popular Word2Vec model is based). Once the method has stored the global co-frequencies of each pair of terms in the corpus, the semantic value of each target word is encoded from the relationship between it and the other terms in the corpus, considering different contexts of interest. More precisely, given two target words i and j, the semantic relationship between them is expressed as the ratio between the probability of co-occurrence of i with a set of so-called context words k (which define some precise contexts) and the probability of co-occurrence of

[6]To explain such linguistic patterns, the authors of the paper mention, for instance, terms like 'hoax', 'fake', 'false', 'fact check', etc.

j with the same set of context words:

$$F(i, j, \tilde{k}) = \frac{P_{ik}}{P_{jk}} \tag{4.8}$$

where $\tilde{k}$ is the set of context words.
In this way, the GloVe methodology is able to calculate the proximity of i and j within various contexts and thus synthesise a machine-readable description of the respective meaning. Thus, after generating the co-occurrence probability matrix, the dense vector is obtained by minimising the following objective function:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = log(X_{ik}) \tag{4.9}$$

where: w_i is the dense embedding vector for the i-th target word to be learned by the model; b_i and $\tilde{b}_k$ are parameters to be selected and X_{ik} is the value for target word i and context word k calculated using the previous formula 4.8. The generation of word embedding, in this case, is therefore a minimization problem in which one wants to make the ratio between the two dense vectors as close as possible to the co-occurrence probability in logarithmic space. Then, in order to avoid evaluating all co-occurrences identically, the method applies a weighting function, reformulating the minimization problem in the form of a linear regression. Finally, since the resolution of this optimization results in two dense vectors - one associated with the target word and the other with the context set -, the last operation consists in summing the two components.

The version of the GloVe model used here is a solution pre-trained on a textual database crawled and provided by the Common Crawl organisation [7]. This version takes into account approximately 2.2 million tokens[8], and for each of them returns a 300-dimensional word-embeddings vector. The same model will be used to run all the English language experiments presented in this thesis. Anyway, since this neural network is pre-trained on a distinct corpus, the internal weights of the embedding layer of CheckerOrSpreader are updated at each iteration of the training process, in order to better adapt to the given linguistic context. Subsequently, the semantic vector representation is passed as input to a Convolutional Neural Network. However, during the training phase, in order to avoid that the updating of the weights within the embedding layer causes an overfitting of users' sample (hence, with the goal of increasing the generalization of the model), the *dropout* technique of regularization is applied. *Dropout* [219] strategy consists of randomly deactivating a fixed portion of neurons in order to decrease the complexity of the model. Then, 32 one-dimensional convolutional filters are applied, with the size of each filter fixed at 4.5 after an optimization process.

[7] Officiel website of Common Crawl organisation: https://commoncrawl.org/the-data/.

[8] The above-mentioned pre-trained version of the GloVe model is available at the following link: https://nlp.stanford.edu/projects/glove/.

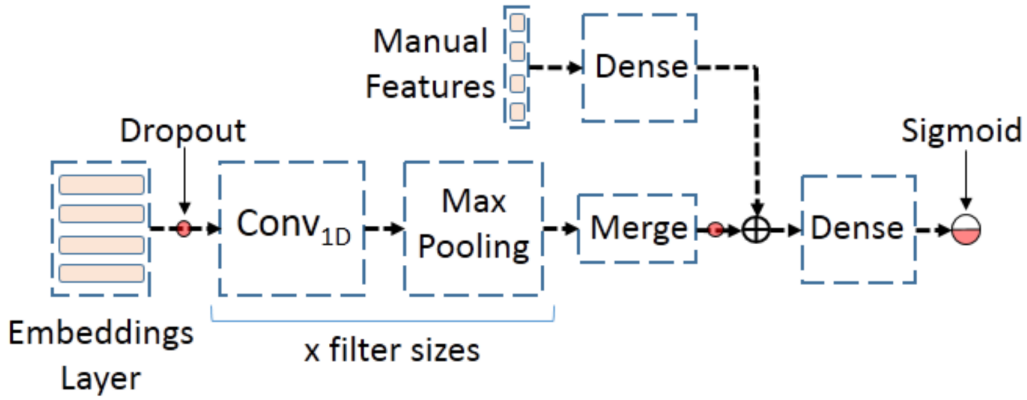

Figure 4-I: Scheme of the CheckerOrSpreader model. Figure retrieved from the original paper by Giachanou et al. [122].

The activation function is the well-known Rectified Linear Unit (ReLU) [220] and the update of the weights is run by the Adaptive moment estimation (Adam) optimiser [221], a version of the stochastic gradient descent algorithm that considers the respective moving averages of both the first two moments of the gradient. Then, the dimension of the convolution output is reduced through a Max Pooling operation, and the resulting vector is concatenated with the product of passing manual features through a Dense neural layer. Finally, what is obtained passes through a final dense layer, which is trained to estimate the weights of each feature, and the prediction is mapped through a sigmoid function. The architecture just described can be seen in Figure 4-I.

The point of interest of the CheckerOrSpreader model lies in the nature of the manual features used to classify users into 'fact checkers' or 'fake news spreaders'. In fact, in addition to the text embedding by GloVe, the architecture is trained for the task of users' profiling with three further classes of features: *(i)* a representation of psycho-linguistic patterns provided by the LIWC software, *(ii)* a codification of personality traits obtained by means of the Five Factor Model, and *(iii)* a set of emotional dimensions extracted with a lexicon-based approach. These categories will be presented in the following Section.

It should also be noted that we trained and tested CheckerOrSpreader model on the two PAN datasets, obtaining an accuracy equal to 0.52 for the English sample and 0.51 for the Spanish sample. This result will be useful for subsequent comparisons.

Chapter 5

Personality Information

This Chapter will explain the methodologies used to extract personality information from the textual content posted by users, which is the basis, inspired by Giachanou et al.'s work [122], for the contribution of the thesis work related to the first goal.

5.1 LIWC Features

The Linguistic Inquiry and Word Count (LIWC) [241] is a software developed for Natural Language Processing tasks, which is able to automatically detect linguistic patterns and map the text into a dense representation composed of 73 psychologically-meaningful linguistic categories. Therefore, this strategy is configured as a lexicon-based method that associates, within a dictionary, a set of predefined classes to several tokens. Using this dictionary, it is then possible to tag the sought psycho-linguistic features and thus obtain evidence of mental and cognitive processes underlying the text of the tweets. The resource used in this project is the LIWC2015 dictionary[1], for both languages (in contrast to what has been done by Giachanou et al., who only tested the approach on a dataset of English users). This tool considers several inflected variants from about 6400 word stems[2], as well as certain selected emoticons; each of these linguistic items has been assigned one or multiple categories among the mentioned 73. These psycho-linguistic classes are arranged in the following hierarchical structure. First of all, three macro-categories have been set up by the creators of the software, which are in turn divided into numerous tags. Below the structure is summarised, and some examples are given:

1. **Linguistic dimensions**, including different types of function words,

[1] Official website of LIWC software at: https://liwc.wpengine.com.

[2] As an example, the stem *hungr** in the dictionary allows for any token that matches the first letters to be counted as a target word of the associated category (in this case, the biological process of ingestion action).

such as pronouns (personal and impersonal), article, prepositions, auxiliary verbs, common adverbs, conjunctions, negations;

2. **Other grammars**, like common verbs (*eat, come, carry*), common adjectives (*free, happy, long*), comparisons (*greater, best, after*), interrogatives (*how, when, what*), numbers, quantifiers (*few, many, much*);

3. **Psychological processes** emerging from the text, which have been grouped by Pennebaker et al. [241] into the following sub-categories:

   3.1. **affective processes**, manifesting negative (*hurt, ugly, nasty*) and positive emotions (*love, nice, sweet*), anxiety (*worried, fearful*), anger (*hate, kill, annoyed*) and sadness (*crying, grief, sad*); this evidence is particularly important for the task of fake news spreaders detection, because authors' involvement is mainly based on the provocation of such mental mechanisms;

   3.2. **social processes**, which relate to primary social structures, such as the family (*daughter, dad, aunt*) and ties of friendship (*buddy, neighbor*), or to the distinction between female (*girl, her, mom*) and male figures (*boy, his, dad*);

   3.3. **cognitive processes**, also crucial for user profiling, as fake news takes advantage of readers' vulnerabilities; within this category are included inner psychological dynamics like insight (*think, know*), causation (*because, effect*), discrepancy (*should, would*), tentative (*maybe, perhaps*), certainty (*always, never*) and differentiation (*hasn't, but, else*);

   3.4. **perceptual processes** concern physical or inward sensations: see (*view, saw, seen*), hear (*listen, hearing*), feel (*feels, touch*);

   3.5. **biological processes** specifically referring to the body (*cheek, hands, spit*), health (*clinic, flu, pill*), sexuality, or the action of ingestion;

   3.6. **drives**, class summarising a range of interactions, impulses and instincts: affiliation (*ally, friend, social*), achievement (*win, success, better*), power (*superior, bully*), reward (*take, prize, benefit*), risk (*danger, doubt*);

   3.7. **time orientations** towards past (*ago, did, talked*), present (*today, is, now*) and future (*may, will, soon*);

   3.8. **relativity** concepts such as motion (*arrive, car, go*), space (*down, in, thin*), time (*end, until, season*);

   3.9. **personal concerns** that generally regulate human life and also end up influencing the opinion on specific issues about, for instance, the political, social and spiritual sphere: work (*job, majors*), leisure (*cook, chat, movie*), home (*kitchen, landlord*), money

(*audit, cash, owe*), religion (*altar, church*), death (*bury, coffin, kill*);

3.10. **informal language**, comprising swear words, netspeak (*btw, thx*), assent (*agree, OK, yes*), nonfluencies (*er, hm, umm*) and fillers (*Imean, youknow*); the effectiveness of an approach based on these precise writing style peculiarities has been demonstrated by various works, including Rashkin et al.'s [13].

In addition, the LIWC resource assigns each category an α parameter about the psychometric reliability, pre-estimated on a corpus composed of about 200000 documents. This values is calculated with two measures:

- Cronbach's uncorrected internal consistency [242]:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_{Y_i}^2}{\sigma_X^2}\right) \tag{5.1}$$

  where k is the number of token within the class, σ_X^2 is the overall variance of the probability that all included tokens actually belong to the given category, and $\sigma_{Y_i}^2$ is the variance of the single token in the sample considered;

- Spearman–Brown's corrected internal consistency [243]:

$$\alpha = \frac{n \cdot \alpha'}{1 + (n-1) \cdot \alpha'} \tag{5.2}$$

  where n is the number of tests carried out to assign the terms extracted from the training corpus to each category and α' is the measure of reliability of current attributions.

In detail, the generation of the LIWC dictionary consisted of the operations explained next. First of all, preliminary conceptual dimensions were defined *a priori* according to external resources - such as standard English dictionaries, thesaurus - like Roget's one [244] - and the Positive Affect Negative Affect Scale (PANAS) by Watson et al. [245]. Then, 26 domain experts individually drew up a list of words for each category. These 26 lists were combined into a unique version by another team of experts based on the relevance of each proposed token. This relevance was measured qualitatively in terms of goodness of fit for each sub-class. Therefore, a word was definitively included in the dictionary if the majority of the annotators agreed on its relevance within a given category. Next, they measured the frequency of each selected token within several corpora offered by multiple online sources (social networks, books, articles, etc.) and those that did not appear at least

once were removed. After filtering, the next step has been the expansion of the dictionary by exploring other linguistic resources: terms that appeared with an high relative frequency for each psycho-linguistic component of the dictionary were treated as new candidates to be included by vote. The search was carried out using Stanford Natural Language Toolkit [246] and Meaning Extraction Helper [247]. The final step has consisted in the psychometric evaluation: considering as response variables the sets of words associated with each category, the two internal consistency statistics were computed. Words that excessively decreased the overall internal consistency of the class were reviewed by the experts and, if necessary, removed after a further majority vote. The whole process just described was repeated several times.

The LIWC machine-readable representation, with which the Machine Learning models will be trained in the various experiments (Chapter 7), coincides with a 73-dimensional vector indicating the total raw occurrence of the LIWC sub-categories within the sub-corpus associated with each author. Giachanou et al. have decided to focus on pronouns, personal concerns, time focus, cognitive and affective processes - like certainty or anxiety -, and informal language. In this paper, on the contrary, all 73 classes are taken into account, as in this way better results were recorded for both the English and the Spanish corpora in terms of accuracy and F-measure. Rashkin et al. [13] also already employed LIWC to detect satire, hoaxes, and propaganda.

5.2 Five Factor Model Features

The second methodology proposed for the extraction of user's personality features from the text is the Five Factor Model [248]. The Five Factor Model is a process of attributing certain psychological characteristics to an individual according to the so-called 'Big Five' taxonomy, developed by Rothmann and Coetzer [249] as a modern evolution of the dispositional approach to the study of human personality and its consequences on behaviour. This theory - also 'trait theory' - stems from the discovery of semantic associations as a result of statistical analysis carried out on a sample of personality survey data. Thanks to these evidences, it was possible to demonstrate that the human psychological dimension can be summarised in only five aspects, referred to by words and expressions recurrent in natural language during the description of the personality of an individual. The five suggested standard factors are listed below (and shown in Figure 5-I):

1. ***Openness to experience***
2. ***Conscientiousness***
3. ***Agreeableness***

4. ***Extraversion***

5. ***Neuroticism (Emotional stability)***

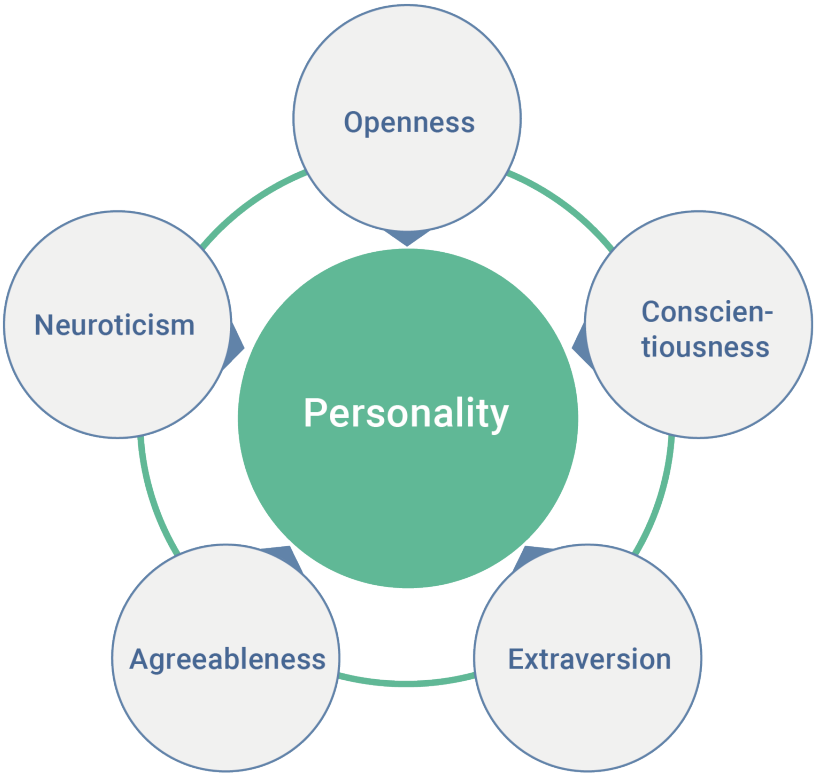

Figure 5-I: The five basic personality factors proposed by the Big Five taxonomy (or *OCEAN* taxonomy).

Openness to experience defines the curiosity of the people towards new ideas and actions outside their belief system, as well as the enjoyment of their intellectual relationship with art, imagination, or novel experiences. This feature is generally associated with a greater creativity, a deeper awareness of feelings, and an increased tendency to develop unconventional opinions. On the contrary, it has been empirically observed that individuals with less openness to experience show a clear pragmatism, a narrower range of interests, and a tendency to adhere more strongly to dogmatic beliefs and traditional or conservative views [250]. When these behavioural trends are analyzed within virtual social platforms, it is reasonable to assume that the scarce presence of this first factor favours the aforementioned phenomenon of informational bubbles (Chapter 2.1), with the consequent formation of closed communities such as echo chambers, the expansion of dangerous narratives

and an easier dissemination of false information. Furthermore, openness is related to the learning mechanism and the absorption of knowledge, other cognitive processes underlying the understanding of the news items and the capacity to distinguish between credible and false articles. From a statistical point of view, openness to experience distributes like a normal variable [251]. The adjectives set as poles of reference for this personality factor are: *inventive* and *curious* - when openness is sufficiently high - and *consistent* and *cautious* - if attraction to new experiences is low.

The so-called *conscientiousness* aggregates the personality traits of sense of duty, control and self-discipline, awareness of one's own psychological characteristics, and willfulness. To sum up, it regulates how the individuals restrain their instincts and whether they act rationally or not. A low level of conscientiousness generally corresponds to a greater flexibility, which can result in weaker and more variable opinions, poorer judgement and, lastly, the lack of reliability. Some studies identify low conscientiousness as the cause of some anti-social behaviour [252], which, in the context of social media, may correspond to some typical inclinations of fake news spreaders such as trolls. Nevertheless, at the same time, too high levels correlate with overconfidence, which turns into closed-mindedness [253] and, as a consequence, vulnerability to fake news. Other research [254] connects an high conscientiousness degree to compulsive behaviour. In this case, the reference adjectives are *efficient* and *organised* - with a sufficient level of conscientiousness - and *extravagant* and *careless* in the opposite case.

According to Matsumoto & Juang [255], *agreeableness* groups the following low-level facets: the level of trust in the others, honesty in communication, altruism, attitude towards interpersonal conflicts, beliefs about oneself, and susceptibility of one's judgement ability to the influence of emotions. In other words, this dimension summarises the nature of the inclinations and the intensity of the reactions of the individuals in their social interactions. The factor, therefore, goes from a high level of concern for social harmony - manifested in altruistic, generous and cooperative behaviour, often linked to optimistic outlooks - to cases where the person shows little empathy, cynicism and scepticism, selfishness and distrust. A very low agreeableness correlates with the Dark Triad, especially with regard to the tendency to prefer manipulation over collaboration [256, 257, 258]. Applying this definition to the study of virtual interactions, it becomes immediately clear the link between this aspect of personality and not only a poor ability to recognise false information, but also the tendency to create it, which - as already seen in the Chapter 2 - can result from a total indifference towards the consequences of such action, or sometimes even from personality disorders such as sadism and psychopathy. Low degrees of agreeableness are associated with the adjectives *critical* and *rational*; conversely, the references are *friendly* and *compassionate*.

Under the definition proposed by Friedman & Schustack [259], *extraver-*

sion refers to the number and frequency of social activities, i.e. the commitment and the attraction towards them. Extroverted individuals are therefore described as active, enthusiastic, collaborative and sometimes dominant in community contexts. It is reasonable to assume that these inclinations can be translated, in the context of virtual interactions within microblogs, into greater visibility and ability to attract other users, even to the role of influencer of the online debate. On the contrary, introverts appear less active and involved in social circumstances, and show a certain independence. According to the 'Big Five' theory, introversion is represented by the adjectives *solitary* and *reserved*. Extroversion, on the other hand, is associated with *outgoing* and *energetic*.

Lastly, Jeronimus et al. [260] define *neuroticism* as an emotional instability that leads to frequent experiences of negative emotions - such as anger, anxiety, frustration, fear and depression -, and which is said to result from a low ability to handle stress or strong external stimuli. As highlighted by Eysenck's theory of personality [261], this excessive emotional reactivity makes high-scorers more prone to interpreting minor issues as insurmountable threats, leading them to a persistent state of alertness. Therefore, this vulnerability reduces rationality and, consequently, logical and decision-making capacities. At this juncture, it is obvious how users with an high score in this personality factor can be more easily manipulated by false information, since - as previously seen - the intent of disinformation content is precisely to induce a state of confusion that, through specific emotional patterns, engages and retains the readers. Numerous investigations [262, 263, 264] have detected a strong correlation between neuroticism and a set of mental and personality disorders: besides anxiety and depression, the studies mention psychosis and paranoia, schizophrenia, bipolarity, dissociative identity and hypochondriasis. In particular, as will be explained later in the Section, the identification of linguistic patterns associated with such psychological pathologies can be an effective prediction tool for the task of fake news spreaders detection. Anyway, low levels of neuroticism, i.e. conditions of emotional stability, are associated with more rational and moderate reactions to external provocations, and a greater resistance to persistent negative feelings.

As mentioned above, the 'Big Five' theory, drawing on numerous studies in Psychology, argues for the existence of semantic associations between a set of words and each of the five factors through which it appears possible to synthesise the personality of an individual. More precisely, this set of words includes the adjectives - or, more generally, the concepts - with which people tend to describe the characteristics that depend on each personality factor (e.g., being creative is often linked to an high openness to new experiences). Exploiting this theoretical framework to extract information about users' personality from their posts means identifying such semantic associa-

tions and mapping the text around the five factors according to the words referring to them. In other words, if a document presents various concepts and adjectives that empirically showed a clear correlation with one of the five dimensions, then that dimension is associated with the text. The core of this methodology lies in the fact that based on the type of semantic association between the terms/concept used by the author and a factor, it is possible to understand how the given individual manifests this basic personality aspect. For example, the adjective *assertive* expresses extroversion, while *indecisive* manifests introversion; an individual described as *organised*, according to the 'Big Five' theory, presents sufficient *conscientiousness*, as opposed to one depicted as *distracted*. Nevertheless, when the authors publish a content in a microblog to express their opinion on a given news item, they have no reason to include adjectives describing themselves in detail, leaving the text devoid of direct references to each of the five factors. Despite this, it is possible to exploit the natural semantic relationship among words in order to connect the text to the five characteristics sought even when these are not mentioned. An effective approach to do this consists in the one proposed by Neuman and Cohen [265]: the evidences of a particular personality trait are summarised into a score, which is calculated as the semantic similarity between the context-free word embedding representations respectively of the text written by the author and of the set of the benchmark adjectives (i.e., the terms empirically observed to be able to encode each of the five personality aspects according to the 'Big Five' framework).

In more detail, for each trait, Neuman and Cohen define a positive and a negative sub-dimension, which corresponds respectively to the possession of a sufficient degree of a given factor or, vice versa, to the evidence of the exact opposite characteristic. For example, the positive dimension of the *agreeableness* trait groups characteristics demonstrating a certain concern for social harmony; conversely, its negative dimension refers to evidences of a disinterest in the quality of social interactions. In addition to these 10 sub-dimensions, they also consider 9 further classes, each relating to a different mental disorder. This is done because, in addition to mapping the text around the five basic factors, they intend to add further information relating to the presence of psychological disorders. This is particularly useful for the fake news spreaders detection task, because, as already seen in the previous sections, the spread of false information is often linked to such severe psychological vulnerabilities. Then, Neuman and Cohen associate a small series of benchmark adjectives to all the 19 sub-dimensions. The associations are shown in the diagram below:

> **Openness** $(+)^3$ $\rightarrow$ *philosophical, abstract, imaginative, curious, reflective, literary, questioning, individualistic, unique, open*

[3] The (+) symbol indicates the positive sub-dimension of the personality trait; (-) points the negative sub-dimension.

Openness $(-) \rightarrow$ *narrow-minded, concrete, ordinary, incurious, thoughtless, ignorant, uneducated, common, conventional, restricted*

Conscientiousness $(+) \rightarrow$ *organized, orderly, tidy, neat, efficient, persistent, systematic, straight, careful, reliable*

Conscientiousness $(-) \rightarrow$ *distracted unreliable incompetent wild inefficient disloyal chaotic confused messy disorganized*

Agreeableness $(+) \rightarrow$ *tender, gentle, soft, kind, affectionate, helpful, sympathetic, friendly*

Agreeableness $(-) \rightarrow$ *cruel, unfriendly, negative, mean, brutal, inconsiderate, insensitive, cold*

Extraversion $(+) \rightarrow$ *dominant, assertive, authoritarian, forceful, assured, confident, firm, persistent*

Extraversion $(-) \rightarrow$ *nervous, modest, quiet, forceless, afraid, shy, calm, indecisive*

Neuroticism $(+) \rightarrow$ *worried, stressed, anxious, nervous, fearful, touchy, guilty, insecure, restless, emotional*

Neuroticism $(-) \rightarrow$ *balanced, stable, confident, fearless, calm, easygoing, relaxed, secure, comforted, peaceful*

Schizoid $\rightarrow$ *indifferent, apathetic, remote, solitary*

Depressive $\rightarrow$ *sad, depressed, hopeless, gloomy, fatalistic*

Avoidant $\rightarrow$ *shy, reflective, embarrassed, anxious*

Dependent $\rightarrow$ *helpless, incapable, passive, immature*

Histrionic → *dramatic, seductive, shallow, hyperactive, vain*

Narcissistic → *selfish, arrogant, grandiose, indifferent*

Compulsive → *restrained, conscientious, respectful, rigid*

Paranoid → *cautious, defensive, distrustful, suspicious*

Schizotypal → *eccentric, alien, bizarre, absent*

In order to calculate the semantic similarity between the text produced by the user and these sets of terms, it is necessary to convert both types of documentes into a machine-readable format, while avoiding an excessive loss of the semantic information expressed by each token. Therefore, Giachanou et al. [122] proposed a word-embedding representation obtained from a model which has been pre-trained on a corpus not linked to a precise thematic context. The solution is thus a Word2Vec[4] architecture pre-trained on the Google News dataset, which returns 300-dimensional vectors for 3 million of words and phrases. This approach will remain unchanged for the experiments on the English dataset; with regard to the Spanish corpus the alternatives will be presented in the Chapter 7. The semantic associations between the tweet text and each sub-dimension is computed by the following process: selected one of the 19 personality traits, the algorithm calculate the cosine similarity between all possible pairs of dense vectors respectively extracted by pre-trained Word2Vec for each term in the original text and for each benchmark adjective included for the given personality trait; then, the result coincides with the average of all the similarity values obtained. The mathematical formulation is as shown below: given the personality trait P, the user's personality score $s(P)$ is calculated as

$$s(P) = \frac{\sum_{i=1}^{n} \sum_{k=1}^{m} \cos(t_i v_{P,k})}{n \cdot m} \tag{5.3}$$

where: t_i is the dense vector representing the i-th token (out of n) composing the sub-corpus of a given author; $v_{P,k}$ is the dense vector representing the k-th benchmark adjective (out of m) included in the set associated with the

[4] A more detailed description of the architecture can be found on the official website at the link: https://code.google.com/archive/p/word2vec/.

personality dimension P; the $n \cdot m$ is the number of all possible combinations between the dense vectors t representing the n terms used by the author and the embeddings v_P of the m adjectives associated with P.

In this way, we obtain a unique vector representation per user which is composed of 19 distinct values quantifying the evidence of a certain dimension - positive or negative - of a personality factor or of a mental disorder: if the value is high, then the text expresses that specific trait in a more intense way.

5.3 Emotional and Additional Features

Inspired by Giachanou et al.'s work [122], in the conducted experiments (Chapter 7) the two aforementioned components of personality information will be combined with two further sets of features:

- ten *emotional dimensions* extracted through the following lexicon-based approach: based on the associations provided by the external resource called NRC Word-Emotion Association Lexicon [157], the final representation corresponds to a 10-dimensional vector counting the global occurrence within user's corpus of 8 specific emotions (*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy*, and *disgust*) and the frequency of tweets with positive or negative sentiment polarity; this type of feature is used to check whether the recurrence of certain emotional levels in the authors' writing style can help classify them as 'fake news spreader' or 'real news spreader';
- *Bag-Of-Words* vectors, i.e. the TF-IDF weights associated to the corpus vocabulary.

Chapter 6

Visual Information

For the reasons presented in the Chapter 2.3.2, this work also aims to assess the effectiveness of visual information for profiling fake news spreaders. In order to do so, it has been necessary to enrich the reference database - i.e., the one provided at the Author Profiling task at PAN 2020 (Chapter 3, Section 3.2.1.1) - with the visual features required for these experiments.

This Chapter explains in detail the contribution of the thesis work regarding the extraction and use of visual information for the task of fake news spreaders detection. The Chapter about the features extraction will also deal with the method implemented to collect images from users' text.

6.1 Image Dataset Creation

Thus, each observation in this dataset - related to each author - has been associated with the set of all possible images that could be extracted directly from the corpus. In other words, the document formed by the union of the last 100 tweets written by each author has been integrated with the visual contents embedded within these same tweets. To achieve this, an algorithm was developed to perform multiple operations. First of all, all URLs were identified among the strings of each user's sub-corpus, defining the search pattern with multiple Regular Expressions. Once a list of web addresses has been assigned to each user, a recursive crawling process is performed on all the lists. In other words, after accessing the resources, the program performs two alternative operations depending on the case: *(i)* if the given URL leads directly to an image, and if this is still available and undamaged, then the algorithm stores it in `JPEG` format within a secondary dataset, associating it with a primary key with which it can be easily traced back to the author by whom it was published; *(ii)* if, on the other hand, the web address does not lead directly to a visual content but to a web page, then the algorithm visits it and tries to retrieve within it elements that present typical image

extensions[1]. This action is performed with a web scraping technique, which simulates human navigation on the World Wide Web using the Hypertext Transfer Protocol (HTTP). In particular, the operation is carried out through an HTML parser[2], which gets HTML tag[3] as a string argument and returns the list of elements within the web page matching with the provided tag. The procedure is recursive because, if the download is not possible, the algorithm moves on to the next image, or directly to the next URL in user's list.

As just seen, in contrast to what is generally done in the literature, the approach presented above does not only store images directly embedded in tweets, but also those belonging to web pages that have been re-shared by the authors. This method is based on the assumption that authors must be considered, in the context of social networks, not only as producers of information, but also as vehicles: when users republish a web page they are supposed to be in turn persuaded by the visual elements it contains, especially when the article has been fabricated specifically to manipulate readers. Therefore, this method provides a broader and deeper description of the visual content associated with the users' virtual history - i.e., the set of the last tweets posted by them.

Finally, since the authors are expected to have republished the same images over the time, in order to avoid unnecessarily overloading the secondary dataset and making the process too computationally demanding, the algorithm removes duplicates within each images set per author. Duplicates are recognised by producing an hash code that uniquely refers to each image file. The method used is the MD5 message-digest algorithm, which receives the content and returns a 128-bit string. With this system, tiny changes give completely different hashes, that can be compared to verify that a copy of the selected image is not already present in the author's set.

The above was obviously applied to both the English and Spanish corpora.

6.2 Visual Feature Extraction

Once we have the set of images published or re-shared by each author from the original dataset, the next step consists in summarising the underlying information into a single representation, which can then be used to assess the existence of visual patterns typical for the category of fake news spreaders. The Chapter 3.1.2.1 has shown the various types of features that can be extracted from images. In this project, it seems more appropriate to implement the category of *semantic* visual features. The reason lies in the fact

[1] In addition to `JPEG`, typical image extensions are: `PNG`, `JFIF`, `TIFF`, `BAT`, etc.

[2] The HTML parsing operation is performed in Python using the popular `Beautiful Soup` library.

[3] For instance, the HTML tag for images is `img`.

that the extraction of both *forensic* and *contextual* features is unnecessarily expensive, and in the fact that the visual *statistical* features appeared less effective than the *semantics* in past research. For the definition of the extraction methodology, the reference point - albeit with clear differences - is Giachanou et al.'s work in [123]. Here, the authors assess the importance of integrating visual information for evaluating the credibility of news items and classify them into 'true' or 'false'. To do so, they mined vectors of features from the images embedded in the selected content, by using pre-trained neural networks. The work presented in this paper only follows a similar methodology of extracting this visual information for each image, but proposes a new perspective for the Author Profiling task: instead of considering the vectors of the images embedding in the tweets to evaluate the credibility of the single content, the new approach offers an average description of all the images posted by each author in the sample, and subsequently evaluates which of the two target classes this description can correspond to. The stages of the visual information extraction process are deployed below. First of all, the images scraped from each user's texts are converted into a machine-readable format. The transformation is performed from an image instance to an array[4] of size $(224, 224)$, by mapping each original pixel to a brightness value between 0 and 255 (more precisely, an 8-bit integer), which is then normalised between 0 and 1 by dividing by 255. After converting the array into a tensor by adding an empty dimension - so that the tensors representing each image can be stacked and processed in batch -, the algorithm performs two pre-processing operations on the tensor: it changes the format from `RGB` to `BGR`, and then it zero-centers each color channel with respect to the original dataset with which the models were pre-trained. The basic structure of a tensor derived from a single image can be seen in Figure 6-I.

Then, the resulting tensors are passed through five different models[5] pre-trained on the popular ImageNet dataset[6]: *(i)* VGG16 [266], *(ii)* VGG19 [266], *(iii)* ResNet50 [267], *(iv)* InceptionV3 [268], *(v)* Xception [269]. The main mechanism behind these architectures, regardless of their reciprocal differences, can be summarised in the concatenation of several convolutional layers which, step by step, are able to extract an increasingly abstract representation of each image and finally synthesise the most useful semantic information for the description of the given input in a machine-readable format. Some characteristics of these models are shown in the Table 6-I. Looking at the table, the Xception architecture shows the best accuracy on the ImageNet validation set, either by measuring performance on a single split ('Top 1 accuracy' column) or by a 5-folds Cross Validation technique

[4]The type of array processed, in this case, in the Python environment is the Numpy array: https://numpy.org/numpy.array.

[5]The mentioned pre-trained models are part of the Keras environment. More details on their applications are available at the link: https://keras.io/api/applications/.

[6]Official website of the ImageNet dataset at http://www.image-net.org.

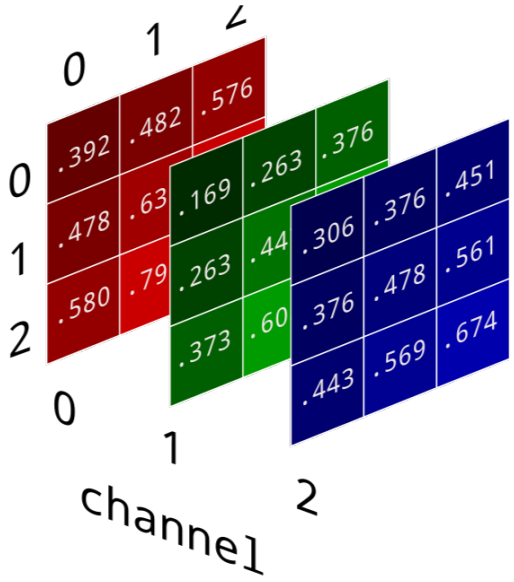

Figure 6-I: Basic structure of a tensor into which an image can be converted for obtaining a machine-readable format. The figure shows separately the three arrays associated with the three Red-Green-Blue channels respectively.

('Top 5 accuracy' column). Despite this, the number of network weights to be estimated is smaller than for all other solutions. From this point of view, the most complex model appears to be VGG19 (which is associated with more than 6 times the number of weights of Xnception).

In order to obtain a feature from these models, instead of a prediction, a popular strategy of Feature Extraction is implemented: each architecture is cut off the last component of classification[7], so that, performing the operation of inference given a three-dimensional array as input, what is obtained in the passage from a convolutional layer to another is a representation. Then, an Average Pooling is applied to the latter in order to synthesize the information into a smaller size. By performing this procedure using each pre-trained model, five vectors of different lengths are obtained for each image: respectively with dimension 512 for VGG16 and VGG19, and 2048 for ResNet50, InceptionV3 and Xception.

Finally, in order to obtain a single visual feature representing, for each of the five extractors, the information from which it is possible to draw the typical characteristics of the visual contents associated with an author, the final step consists in an averaging operation of the vectors obtained from all the images posted or re-shared by the same author. More precisely, given n vectors of the same length, the result of an averaging operation among

[7]The Keras environment offers the possibility of dismantling deep architectures into modules. In this case, removing the classifier means cutting the last module.

Table 6-I: Preliminary insight into the models pre-trained on ImageNet and used for the extraction of semantic visual features: the two accuracies 'Top-1' and 'Top-5' refer to the best result obtained after evaluating the performance on the ImageNet validation set through Cross Validation with one fold and 5 folds respectively; 'Parameters' is the number of weights to be estimated to define the model.

Model	Top-1 Accuracy	Top-5 Accuracy	Parameters
VGG16	0.713	0.901	138357544
VGG19	0.713	0.900	**143667240**
ResNet50	0.749	0.921	25636712
InceptionV3	0.779	0.937	23851784
Xception	**0.790**	**0.945**	22910480

them consists of a new vector comprising the averages of all the n values at the same position within the vector, calculated for each position. In case no images were available for a particular author, vectors of zero have been assigned.

It is worth mentioning that another representation for visual features has been experimented with the Local Binary Pattern [166]. However, this statistics solution was omitted as it offered poor results in all the experiments.

Chapter 7

Experiments and Results

This chapter is organized in order to show the results of the experiments conducted on both languages and related to the three project goals: *(1)* assessing the effectiveness of psycho-linguistic features; *(2)* evaluate the effectiveness of the visual semantic features; *(3)* verify whether the introduction of visual and personality information can improve the performances of state-of-the-art models (Buda-Bolonyai's and Pizarro's solutions), in the special case of profiling fake news spreaders.

7.1 Experimental Setup

In order to make valid comparisons, all the experiments have been performed on the same training set and test set provided at the Author Profiling task at PAN 2020 - composed of English and Spanish corpora extracted from Twitter for 500 authors per language (as explained in the Chapter 3.2.1.1) -, subsequently enriched with the images published or re-shared by the same authors (with the procedure outlined in the Chapter 6.1).

The metric used to assess the performances of the models is the same as that taken into account at the Author Profiling task at PAN 2020: the accuracy, calculated with the formula below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.1}$$

where TP and TN are respectively the number of True Positive and True Negative, while FP and FN are the number of False Positive and False Negative recorded out of the assignments returned by the model on the test set. As the work involves two separate corpora written in two different languages, the overall performance of each solution is calculated as the average of the two results obtained. Since the two target classes within the corpora ('fake news spreader' and 'real news spreader') are balanced, the additional metric of F-measure, although computed during the tests, is omitted from the discussion because it is redundant.

The entire project was developed within the Python programming environment. The main libraries used are Natural Language Toolkit `ntlk`[1] for text preparation and processing, Scikit-learn `sklearn`[2] for the application of the key Machine Learning techniques, Hyperopt `hyperopt`[3] for the optimisation of parameters and hyper-parameters.

7.2 Effectiveness of Psycho-Linguistic Features

As a first objective, the project aims at verifying the potential, for the author profiling task, of the two psycho-linguistic components: *(i)* the 73-dimensional feature describing lexical patterns and mental processes in the text, extracted through the LIWC software; *(ii)* the 19-dimensional representation about personality traits and mental disorder evidences, provided by the application of the Five Factor Model. To do so, we tested all the possible combinations among these two feature sets *(i)*-*(ii)* - separately or jointly considered - with three additional components: *(iii)* the emotional dimensions tagged through a lexicon-based Sentiment Analysis (already presented in the Chapter 5.3), *(iv)* the vectors resulting from the calculation of TF-IDF weights in a Bag-Of-Words approach (also mentioned in Chapter 5.3), and, lastly, *(v)* the statistical stylistical features implemented within Buda-Bolonyai's solution (Chapter 4.1.2). The experiments introduce feature sets not strictly related to the psychological aspects of the author (in particular, *(iv)* and *(v)*), in order to test not only the individual performances of *(i)* and *(ii)*, but also whether their predictive ability increases with the addition of emotional and Bag-Of-Words representations, as inspired by Giachanou et al. [122].

In order to obtain the LIWC features on the Spanish text, a specific version of the LIWC2015 software was used. As for the extraction of personality traits and mental diseases signals defined by the Five Factor Model, on the other hand, we performed a manual translation of the benchmark adjectives set for each of the 19 classes. The output of this translation can be seen in the following diagram:

> **Openness** $(+)^4 \rightarrow$ *filosófico, abstracto, imaginativo, curioso, reflexivo, culto, inquisitivo, independiente, único, abierto*

[1] The official documentation of Natural Language Toolkit library is available at https://www.nltk.org.

[2] The official documentation of Scikit-learn library is available at https://scikit-learn.org/stable/.

[3] The official documentation of Hyperopt library is available at https://github.com/hyperopt/hyperopt.

[4] As seen previously, the (+) symbol indicates the positive sub-dimension of the personality trait; (-) points the negative sub-dimension.

Openness (−) → *cerrado, concreto, común, desinteresado, desconsiderado, ignorante, deseducado, ordinario, convencional, limitado*

Conscientiousness (+) → *organizado, ordenado, arreglado, aseado, eficiente, persistente, metódico, directo, cuidadoso, fiable*

Conscientiousness (−) → *distraído, faltoso, incapaz, salvaje, ineficiente, traicionero, caótico, confundido, desordenado, desorganizado*

Agreeableness (+) → *tierno, amable, blando, atento, cariñoso, servicial, solidario, amigable*

Agreeableness (−) → *cruel, antipático, negativo, malo, brutal, desconsiderado, insensible, frío*

Extraversion (+) → *dominante, resuelto, autoritario, enérgico, seguro, confiado, firme, persistente*

Extraversion (−) → *nervioso, modesto, tranquilo, flojo, preocupado, tímido, calmado, indeciso*

Neuroticism (+) → *preocupado, estresado, ansioso, nervioso, miedoso, susceptible, culpable, inseguro, inquieto, sensible*

Neuroticism (−) → *equilibrado, estable, confiado, valiente, calmado, suelto, relajado, seguro, consolado, tranquilo*

Schizoid → *indiferente, apático, distante, solitario*

Depressive → *triste, deprimido, desesperado, pesimista, fatalista*

Avoidant → *tímido, reflexivo, avergonzado, inquieto*

Dependent → *indefenso, incapaz, pasivo, inmaduro*

Histrionic → *dramático, atractivo, superficial, hiperactivo, vanidoso*

Narcissistic → *egoísta, arrogante, exagerado, indiferente*

Compulsive → *contenido, meticuloso, respetuoso, rígido*

Paranoid → *prudente, defensivo, desconfiado, receloso*

Schizotypal → *excéntrico, extraño, raro, distraído*

In the same context, the embedding used to convert these sets and the input text into a machine-readable format comes from a FastText neural network [270] based on a Skip-gram approach, pre-trained on Spanish Unannotated Corpora[5] and considering 1.3 millions of Spanish tokens. The choice of this strategy seems reasonable, since this model is pre-trained on documents that do not come from a specific context, and therefore does not overfit any semantic domain.

The predictive architectures employed are of three different types and depths: a Logistic Regression (LogReg), a Convolutional Neural Network (CNN), and a Long Short-Term Memory (LSTM). As far as Logistic Regression is concerned, its characteristics and advantages (as, above all, its ability to handle multicollinear variables in the training matrix given as input) have already been discussed in the Chapter 4.1.1. With regards to the use of a Convolution Neural Network model, the solution implemented remains the one on which the CheckerOrSpreader model is based: the passage, with the application of a *Dropout* operation, of an input embedding to a layer performing a sequence of one-dimensional convolutions, whose result is down-sized before being merged with the additional manual features (among the five just listed) into a final representation on the basis of which a last dense layer learns to carry out the classification into the 'fake news spreader' or 'real news spreader' classes (Figure 4-I). The third predictive architecture tested is exactly the same as that just explained for the CheckerOrSpreader model, the only difference being that it replaces the Convolutional Neural

[5]More details on the Spanish Unannotated Corpora are available at the link: https://github.com/josecannete/spanish-corpora.

Network with a Long Short-Term Memory model. The latter [271] is configured as a complex evolution of Recurrent Neural Networks (RNN), capable of effectively modelling multiple dependencies between words within the corpus, even when these words are separated by a considerable number of other terms. This mechanism of modelling the relationships among tokens in long documents (just like in the case of our sub-corpora, derived from the last 100 tweets of each user) is implemented through 4 types of blocks: a *basic cell*, an *input gate*, a *forget gate* and an *output gate*. The three gates are equivalent to sequences of matrix operations, whose weights sets are estimated during the training process. Their function is to regulate the flow of data in and out of the cell, i.e. the flow of information with which the recurrent neural network is trained in order to optimise the result but at the same time avoid overfitting the training set. Furthermore, the model functioning solves the typical problem of vanishing gradients, i.e. it reduces the risk that, during the process of updating the weights through the back-propagation, the value of the gradients tends to zero, or to infinity. To do this, the first operation is to identify the redundant information, which will be omitted in the next iteration. This is done by the *forget gate*, which corresponds to a sigmoid function σ that receives two kinds of data: the output of the previous unit h_{t-1} at time $t-1$ and the current input at time t (Figure 7-I). The output of

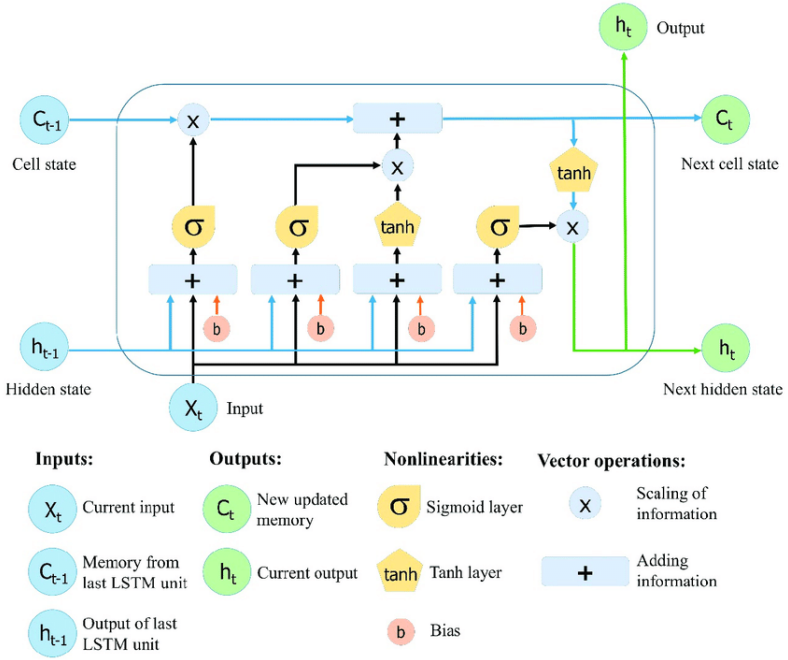

Figure 7-I: Basic structure of a Long Short-Term Memory model, as schematised in [272].

the *forget gate* at the current time f_t is, thus, derived by applying a sigmoid

function to the sum of the *forget gate* bias b_f and the *forget gate* weights matrix W_f:

$$f_t = \sigma(W_f[h_{t1}, X_t] + b_f) \tag{7.2}$$

The next step consists in storing the information from the current input X_t to update the cell state (that is making the switch from C_{t-1} to C_t). This operation involves a second sigmoid function and a *tanh* function. The first filters the information for the next iteration - estimating a zero weight if the data is to be omitted and 1 in the opposite case. The *tanh* function, on the other hand, applies an additional importance weight (between -1 and 1) on the information that is not omitted by the previous sigmoid. In the final step, the output value is computed as the filtered version of the output cell state O_t: the multiplication of a binary value returned by a σ and the importance value (between -1 and 1) produced by the *tanh* function.

As said, the CNN-based and LSTM-based architectures consider as input the vector representation of the tweets provided by a pre-trained model. For the English sample, as explained in the Chapter 4.3, this latter corresponds to a Global Vectors for Word Representation (GloVe) solution pre-trained on the Common Crawl dataset. In the experiments on the Spanish corpus, it was decided to use the same FastText method [270] implemented for computing the dense vectors during the extraction of the Five Factor Model features. Both models are re-trained during the training phase to better learn the semantic relationships within the specific corpus.

The best results in terms of accuracy for each combination among the three classifiers and the various features sets are reported in Table 7-I and Table 7-II. The best solution for both languages is a Logistic Regression

Table 7-I: Best combinations among LogReg, CNN and LSTM and psycholinguistic, emotional (Emo), BOW and Buda-Bolonyai's (Stat.) features on English text.

Model	Personality Feat.	Other Features	Accuracy
LogReg	FFM	BOW	**0.69**
LogReg	FFM, LIWC	BOW, Emo	0.69
LogReg	FFM, LIWC	BOW, Emo, Stat.	0.64
LogReg	FFM, LIWC	-	0.57
LogReg	FFM	-	0.55
LSTM	FFM, LIWC	BOW	0.54
CNN	LIWC	-	0.52

trained on the mix of personality scores and TF-IDF values, offering an ac-

Table 7-II: Best combinations among LogReg, CNN and LSTM and psycho-linguistic, emotional (Emo), BOW and Buda-Bolonyai's (Stat.) features on Spanish text.

Model	Personality Feat.	Other Features	Accuracy
LogReg	FFM	BOW	**0.75**
LogReg	FFM, LIWC	BOW, Emo	0.74
LogReg	FFM, LIWC	BOW, Emo, Stat.	0.73
LogReg	FFM, LIWC	BOW	0.72
LogReg	LIWC	BOW	0.72
LogReg	FFM, LIWC	-	0.7
LogReg	LIWC	-	0.7
LogReg	FFM	-	0.61
CNN	FFM, LIWC	BOW, Emo	0.53

curacy of 0.69 for English and 0.75 for Spanish. In both cases the value is lower with respect to the best performances respectively by Buda-Bolonyai (0.75 on the English corpus) and Pizarro (0.82 on the Spanish corpus). However, while in the Spanish case this difference with respect to Pizarro's result is statistically significant with a confidence level set at 95%, as regards English it is not. Therefore, the significance data about these differences confirms that personality scores derived by Five Factor Model - in combination with a Bag-Of-Words approach - are powerful enough to significantly conform the state-of-the-art performances on the author profiling task only in the case of English text. Furthermore, on the Spanish corpus, training a Logistic Regression only on the TF-IDF values provides even a greater accuracy than any other combination with the psycho-linguistic features (0.76), further demonstrating their lesser effectiveness.

On both samples, the personality scores without Bag-Of-Words vectors offer a poorer result: 0.55 for English and 0.61 for Spanish. In the first case, the addition of LIWC representation slightly increases the result (0.57), but when considered alone it offers the worst performance (0.52). In contrast, the only LIWC representation on the Spanish authors allows to reach a better accuracy with respect to the personality scores alone.

Looking at Table 7-I, it is possible to conclude that emotional and Buda-Bolonyai's features in combination with the personality information make a little contribution to the accuracy result.

Finally, it is important to note that deeper models like CNN or LSTM always give worse results than Logistic Regression. The cause could be found in the fact that this latter function is able to intrinsically manage the strong collinearity between variables in a better way than the two others architectures, which is the condition occurring among the psycho-linguistic features.

In addition, the limited size of both datasets may penalise more complex models, which in general require more data.

7.3 Effectiveness of Visual Features

The evaluation of the usefulness of an author's 'visual profile' for fake news spreaders detection has involved the experimentation of all possible combinations among the five vector representations obtained from each truncated neural network respectively, and, at a later time, mixing also with the psycholinguistic components, as well as emotional dimensions. The tests have been carried out by training a simple Logistic Regression, since as mentioned in the Chapter 7.2 this ensured an efficient management of the very strong multicollinearity among the visual variables. Tables 7-III and 7-IV present the best results for each type of combination on both languages.

Table 7-III: Best results for all types of combination among visual features alone and psycho-emotional information (English).

Visual Feat.	Personality Feat.	Accuracy
VGG19, RN, XNC	LIWC	**0.675**
VGG19, RN, XNC	FFM, LIWC	0.665
VGG16, RN, XNC	FFM, LIWC, Emo	0.66
VGG16,VGG19, XNC	FFM	0.655
VGG16, XNC	-	0.59

Table 7-IV: Best results for all types of combination among visual features alone and psycho-emotional information (Spanish).

Visual Feat.	Personality Feat.	Accuracy
VGG16	FFM, LIWC	**0.706**
VGG16	LIWC	**0.706**
VGG16	FFM, LIWC, Emo	0.7
VGG16	FFM	0.68
VGG16	-	0.553

Observing them, we can see that in both cases the best solution remains the union of the linguistic patterns extracted from the LIWC with visual information, even if the results on the two corpora (0.675 for English and 0.706 for Spanish) are statistically significantly lower than the best perfor-

mances of the PAN task. Furthermore, personality scores appear always less suitable for mixing with visual information.

It is important to note that, while in the first case the features set that best combines the psycho-linguistic features is composed of three different representations - VGG19, ResNet and Xnception -, in the second one VGG16 vector appears to be the only one useful.

In general, visual information alone offers an accuracy that may seem poor: for English authors the combination VGG16-Xnception achieves 0.59, and for Spanish the VGG16 vector gives 0.553. Actually, it is worth mentioning that these solutions, although they do not consider any textual information at all, still manage to outperform the result achieved by the original CheckerOrSpreader model on the same PAN test sets (0.52 for English and 0.51 for Spanish). In the first case, this difference is even statistically significant - with a 95% confidence level.

7.4 Improvements to State-of-the-art Models

The possibility of improving the performances of state-of-the-art systems has been explored by replacing and/or adding the visual and personality features to the sets originally used by the respective authors. To reduce the computational weight and the training time during the experiments, only the best performing combinations between visual and psycho-emotional information have been implemented in the new models (those offering the best accuracies in the tests discussed by the previous Chapter 7.3). In the following sections, the results derived from the changes to the two best systems at PAN 2020 will be illustrated separately.

7.4.1 Variations to Buda & Boloyai's

This chapter will explain the alterations made with the goal of testing the effectiveness of introducing visual and personality information into the solution associated with the best score on the English sample at the Author Profiling task at PAN 2020. For the structure of Buda & Bolonya's system, these modifications essentially concerned the only input matrix of the fifth baseline. In fact, while we maintained unchanged the simultaneous training of the other four baselines (Logistic Regression, Support Vector Machine, Random Forest and the first XGBoost) on the weighted word n-grams, we have tried out several combinations of features within the set on which the second XGBoost algorithm is trained. In addition, instead of applying a Logistic Regression as the only ensemble method, at each test also the Linear Regression has been experimented and its results have been compared. In this case, with the aim of producing more conservative and reliable models, a penalty derived from the Euclidean norm L_2 is applied to the estimation of the regression coefficients, or the degrees of 'importance' of each baseline. More precisely, this

ensemble method coincides with the definition of Ridge Regression. As will be noted, the choice of this strategy has proved to be useful in order to build generally more accurate classifiers. In order to optimise the parameters and hyperparameters of both the five sub-classifiers and the ensemble model, an optimisation process is implemented using the Cross Validation technique, with 10 folds. In particular, with regard to the parameter α which drives the intensity of the regularisation during the Ridge Regression, the algorithm performs a grid search in the interval $\{1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 15\}$. Also for the ensemble Logistic Regression the estimation of the weights is adjusted by a *shrinkage* technique, but in this case the norm applied is L_1. Here again, the hyperparameter that sets the strength of the regularization is optimized with a grid search.

In Table 7-V - focused on the English dataset -, we can see that any replacement and integration of visual/personality information in the ensemble model improves the original result, with a maximum of 0.775 - which, however, is not statistically significantly higher with respect to the original accuracy of 0.75, corresponding to the combination of different baselines trained on N-grams plus an XGBoost model fed with personality scores and VGG16-Xnception vectors. The only exception - an accuracy worse than the original one - is found when only integrating the Five Factor Model representation.

Table 7-V: Results from variations on Buda-Bolonyai's systems, by changing the features set on which the XGBoost algorithm in trained and experimenting also Linear Regression as ensemble method (English). The grey line points out the original score. Combinations including 'Emo' have been omitted because its integration is irrelevant.

Ensemble	Features	Accuracy
LinReg	*Ngrams*+ FFM+VGG16,XNC	0.775
LinReg	*Ngrams*+ VGG16,VGG19,RN,INC,XNC	0.77
LinReg	*Ngrams*+ FFM,LIWC	0.77
LinReg	*Ngrams*+ FFM,LIWC+VGG19,RN,XNC	0.765
LinReg	*Ngrams*+ FFM,LIWC+Stat.	0.765
LogReg	*Ngrams*+ FFM+Stat.	0.765
LogReg	*Ngrams*+ Stat.	0.75
LogReg	*Ngrams*+ FFM	0.725

In Table 7-VI, the opposite is observed: on Spanish authors, any modification worsens the original result.

Table 7-VI: Results from variations on Buda-Bolonyai's systems, by changing the features set on which the XGBoost algorithm in trained and experimenting also Linear Regression as ensemble method (Spanish). The grey line points out the original score.

Ensemble	Features	Accuracy
LinReg	*Ngrams*+ Stat.	0.805
LinReg	*Ngrams*+ FFM,LIWC	0.802
LinReg	*Ngrams*+ VGG16	0.797
LinReg	*Ngrams*+ FFM,LIWC+ Emo	0.797
LinReg	*Ngrams*+ FFM+VGG16	0.797
LinReg	*Ngrams*+ FFM,LIWC+Emo+Stat.	0.792
LogReg	*Ngrams*+ FFM,LIWC+VGG16+Emo	0.792
LogReg	*Ngrams*+ FFM,LIWC+Stat.	0.792
LogReg	*Ngrams*+ FFM,LIWC+VGG16	0.792
LinReg	*Ngrams*+ FFM	0.792
LinReg	*Ngrams*+ FFM+Stat.	0.792

7.4.2 Variations to Pizzaro's

This second section will show the principle variations realized on the system with the best performance on the Spanish corpus at the Author Profiling task at PAN 2020. Since the original model was iteratively trained on only several mixtures of n-grams extracted from pre-processed text, the changes on Pizarro's system consisted in simple concatenations to the TF-IDF weights originally considered of the new features set, including: *(i)* the psycho-emotional representations (from the LIWC software, the Five Factor Model and the Sentiment Analysis with the NRC lexicon), *(ii)* the visual information (extracted by the five pre-trained and truncated neural networks), and *(iii)* 17 Buda-Bolonyai's stylistic statistics.

As this system is based on the search for the optimal pre-processing operations, for conducting correct experiments it is advisable to estimate the personality scores only once by replicating the same original text preparation performed by Giachanou et al. [122]. The Five Factor Model representation is, in fact, affected by the pre-processing step much more than the other features used: the model must clean and compress the author's text in the most effective way to calculate the correct similarity with the low-dimensional vectors of the benchmark adjectives. For this reason, variations in text preparation could penalize the usefulness of these features, as well as make it impossible to judge their performance correctly.

Integrations almost always lead to a performance worsening, as far as

Table 7-VII: Results from variations on Pizarro's systems, by concatenating the new features sets (English). The grey line points out the original score. Combinations including 'Stat.' have been omitted because its integration is irrelevant.

N-grams Feat.	Other Feat.	Accuracy
word	FFM	0.76
word-char	-	0.735
word	VGG19, RN, XNC	0.715
word-char	FFM, LIWC	0.715
word	FFM + VGG16, XNC	0.715
word	FFM, LIWC + VGG19, RN, XNC	0.705
word	FFM, LIWC + Emo	0.7

both languages are concerned (Tables 7-VII and 7-VIII). The exception on the English corpus occurs with the addition of personality scores: in this case, the accuracy increases from the original 0.735 to 0.76; an improvement, however, not yet statistically significant at 95%. On Spanish-writing authors (Table 7-VIII), the only improvement in the accuracy is due to the concatenation of the VGG16 vector, or equally of the union between both Five Factor Model and VGG16 representations: for both solutions, the result rises from 0.82 to 0.832. Nevertheless, the performance continues not to be superior in a statistically significant way.

A final insight must be given about the optimal pre-processing pipelines. As for the best solution on the English sample, given by the union between the Five Factor Model features and word n-gram, the selected text preparation operations are those in the set ***v1_1***: reduce the number of repeated characters in a word (*reduce_length*) and replace emojis with word representation (*demojify*). The optimal window selected by the algorithm extracts uni-grams and bi-grams, removing those with a frequency of less than 2% and more than 85%. On the other hand, when profiling authors writing in Spanish, the best approach - joining personality scores, vectors extracted by the VGG16 network and both word and character n-gram - is associated with the two following pre-processing pipelines:

- for word n-gram, ***v3_0***, which consists in the sequence of the downcasing of all the letters (*downcase*), reduce the number of repeated characters (*reduce_length*), emojis conversion (*demojify*) and the replacement of the anonymized tags referred to URLs, users' name and hashtags; again, uni-grams and bi-grams are mined, but in an occurrence range between 4% and 70%;

Table 7-VIII: Results from variations on Pizarro's systems, by concatenating the new features sets (Spanish). The grey line points out the original score. Combinations including 'Stat.' have been omitted because its integration is irrelevant.

N-grams Feat.	Other Feat.	Accuracy
word-char	FFM + VGG16	0.832
word-char	VGG16	0.832
word-char	-	0.82
word-char	FFM	0.812
word-char	FFM, LIWC + VGG16	0.772
word-char	FFM, LIWC+VGG+Emo	0.751
word-char	FFM, LIWC	0.72
word-char	FFM, LIWC + Emo	0.716

- for character n-gram, ***v0_ 1***, which only includes the replacement of redundant characters (*reduce_ length*); here, the order of the n-grams goes from 1 to 5, and the frequency limits are 6% and 75%.

Chapter 8

Conclusions

Online disinformation and misinformation are expanding phenomena, recently increasingly debated due to the obvious and dramatic influences they have not only on the discussions within the social media, but also on the social structures, and on the political and even economic dynamics. For this reason, it is of primary importance that scientific research is concerned with defining these issues in all their forms and countering the spread of false information, especially in environments such as social media, where debunking actions by domain experts are slowed down or insufficient. In this context, as we have seen, the computational branch can offer great help by providing methods that can reliably and automatically assess the credibility of content. It is therefore possible to conduct analyses on various levels of the text - lexical, logical and semantic - as well as on different elements of the content itself - visual, textual, structural, psychological and linked to the social media of dissemination.

A particularly useful approach is the assessment of user credibility, in order to identify individuals who have a tendency to create and share fake news within the virtual community: the so-called fake news spreaders. In this project, this perspective of author profiling has been addressed through several categories of features: word embedding (provided by the pre-trained models FastText and GloVe), Bag-Of-Words, emotional dimensions (resulting from a lexicon-based analysis) and, in particular, psycho-linguistic patterns (extracted by LIWC software), personality traits and evidences of mental disorders (derived applying the Five Factor Model), and semantic visual information. By experimenting with numerous combinations of these features, this paper has offered an insight into the predictive capabilities of each, as well as their ability to improve the performance of state-of-the-art models.

It is reasonable to summarise what has been observed in the tests by looking at the Tables 8-I and 8-II, which offer an insight on the best solutions - in terms of model accuracy - among all those tested in the experiments

presented in the previous Sections 7.2, 7.3 and 7.4.

Table 8-I: Best overall solutions for English corpus.

Combination	Model	Features	Accuracy
TXT+PERS+IMG	LinReg Ensemble	N-grams + FFM + VGG16, XNC	**0.775**
TXT + IMG	LinReg Ensemble	N-grams + VGG16,VGG19,RN,INC,XNC	0.77
TXT+PERS	LogReg Ensemble	N-grams + FFM, LIWC	0.77
TXT+PERS+STAT	LogReg Ensemble	N-grams + FFM, LIWC + Stat.	0.765
TXT+PERS	Linear SVC	N-grams + FFM	0.76
TXT+STAT	*Buda-Bolonyai's*	N-grams + Stat.	0.75
TXT	*Pizarro's*	N-grams	0.735

Table 8-II: Best overall solutions for Spanish corpus.

Combination	Model	Features	Accuracy
TXT+PERS+IMG	Linear SVC	N-grams + FFM + VGG16	**0.832**
TXT + IMG	Linear SVC	N-grams + VGG16	0.832
TXT	*Pizarro's*	N-grams	0.82
TXT + PERS	Linear SVC	N-grams + FFM	0.812
TXT + STAT	*Buda-Bolonyai's*	N-grams + Stat.	0.805

First, it appears that the task of profiling fake news spreaders can be addressed more effectively with a combination of N-grams, personality information and visual features, both on English and Spanish corpora. In particular, we observe that the most powerful psycho-linguistic feature in both cases is offered by the Five Factor Model, but the union with the LIWC linguistic patterns often penalizes the result. Therefore, these results demonstrate the relevance of user's visual and personality information for an efficient profiling.

In both corpora, the second best solution combines visual information with textual features. A second consideration can thus be made on the effectiveness of the introduction of visual features: although they offer lower results if alone, when combined with N-grams they always offer better performance with respect to the mix of text and personality information. In general, even in combination with psycho-linguistic features, visual information offers good results. This answers the research question indirectly posed by the second goal of this project.

As regards the first objective, personality scores, modeled together with BOW by a Logistic Regression, are able to offer a result statistically not

inferior to state-of-the-art solutions only in the English case, while in the Spanish case the accuracy approaches the value reached by Pizarro without falling within the 95% confidence interval. In this context, it was noted that it is advisable to use less complex models.

Finally, with regards to the third goal of the project, it is possible to conclude that the integration of visual/personality information allows to improve the performances of state-of-the-art models in many cases, although to a not statistically significant extent.

Bibliography

[1] Allcott H., Gentzkow M.: Social media and fake news in the 2016 election. In: National Bureau of Economic Research (2017)

[2] Zannettou S., Sirivianos M., Blackburn J., Kourtellis N.: The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. In: Journal of Data and Information Quality (JDIQ), 11(3):1–37 (2019)

[3] Robinson-Riegler B., Robinson-Riegler, G.: Cognitive Psychology: Applying the Science of the Mind. Allyn Bacon, 313 (2004)

[4] Loftus E. F.: Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. In: Learning Memory 12(4), 361–366 (2005)

[5] Loftus E. F., Wylie L. E., Patihis L., McCuller L. L., Davis D., Brank E. M., Bornstein B. H.: Misinformation effects in older versus younger adults: A meta-analysis and review. In: The Elderly Eyewitness in Court (2014)

[6] Briggs-Myers, I., Myers, P.B.: Gifts differing: Understanding Personality Type. Davies-Black Publishing. (1995)

[7] Lee K.: Age, Neuropsychological, and Social Cognitive Measures as Predictors of Individual Differences in Susceptibility to the Misinformation Effect. In: Applied Cognitive Psychology 18(8), 997–1019 (2004)

[8] Ward R.A., Loftus E.F.: Eyewitness performance in different psychological types. In: Journal of General Psychology 112(2), 191–200 (1985)

[9] Zajonc R. B.: Attitudinal effects of mere exposure. In: Journal of personality and social psychology 9(22):1 (1968)

[10] Campbell W. J.: Yellow journalism: Puncturing the myths, defining the legacies. Praeger (2001)

[11] Chen Y., Conroy N. J., Rubin V. L.: Misleading online content: Recognizing clickbait as false news. In: MDD (2015)

[12] Politifact: The more outrageous, the better: How clickbait ads make money for fake news sites. http://www.politifact.com/more-outrageous-better-how-clickbait-ads-make-mone/ (2017).

[13] Rashkin H., Choi E., Jang J. Y., Volkova S., Choi Y.: Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2931–2937 (2017)

[14] Situngkir H.: Spread of hoax in social media. In: MPRA Paper 30674 (2011)

[15] Oh O., Kwon K. H., Rao H. R.: An exploration of social media in extreme events: Rumor theory and twitter during the haiti earthquake. In: ICIS (2010)

[16] Parlett M. A.: Demonizing a President: The 'Foreignization' of Barack Obama. Praeger (2014)

[17] Zubiaga A., Liakata M., Procter R. , Hoi G.W.S., Tolmie P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. In: PloS one (2016)

[18] Kwon S., Cha M., Jung K., Chen W., Wang Y.: Aspects of rumor spreading on a microblog network. In: SocInfo (2013)

[19] Sánchez-Junquera J., Rosso P., Manuel Montes-y-Gómez, Ponzetto S.: Unmasking Bias in News. arXiv:1906.04836 (2019)

[20] Royal Society: COVID-19 vaccine deployment: Behaviour, ethics, misinformation and policy strategies. https://royalsociety.org/set-c-vaccine-deployment.pdf (2020)

[21] Selby J.: The Trump presidency, climate change, and the prospect of a disorderly energy transition. In: Review of International Studies 45(3), 471-490. (2019)

[22] Tharoor I: Bolsonaro, Trump and the nationalists ignoring climate disaster. In: Washington Post. https://www.washingtonpost.com/bolsonaro-trump-ignoring-climate-disaster/ (2019)

[23] Kogan S., Moskowitz T. J., Niessner M.: Fake News: Evidence from Financial Markets. In: SSRN (2020)

[24] Confessore N.: Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. In: The New York Times. https://www.nytimes.com/cambridge-analytica-scandal.html (2018)

[25] Meredith S.: Facebook-Cambridge Analytica: A timeline of the data hijacking scandal. In: CNBC. https://www.cnbc.com/facebook-cambridge-analytica.html (2018)

[26] Kozlowska H.: The Cambridge Analytica scandal affected 87 million people, Facebook says. In: Quartz (2018)

[27] Lewis P., Hilder P.: Leaked: Cambridge Analytica's blueprint for Trump victory. In: The Guardian (2018)

[28] Field M., Wright M.: Russian trolls sent thousands of pro-Leave messages on day of Brexit referendum, Twitter data reveals. In: The Telegraph (2018)

[29] 'Enabling further research of information operations on Twitter' In: Twitter Blog (2019)

[30] Baraniuk C.: Beware the Brexit bots: The Twitter spam out to swing your vote. In: New Scientist (2016)

[31] Howard P. N., Kollanyi B.: Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. arxiv:1606.06356 (2016)

[32] Mostrous A., Bridge M., Gibbons K.: Russia used Twitter bots and trolls 'to disrupt' Brexit vote. In: The Times (2017)

[33] Russian Twitter accounts promoted Brexit ahead of EU referendum – Times newspaper. In: Reuters (2017)

[34] Castle S., Landler M.: No One' Protected British Democracy From Russia, U.K. Report Concludes. In: The New York Times. (2020)

[35] Digital, Culture, Media and Sport Committee: Disinformation and 'fake news': Interim Report. https://publications.parliament.uk/pa/cm201719/ (2018)

[36] Bobba G.: Social media populism: features and 'likeability' of Lega Nord communication on Facebook. In: European Political Science 18(6) (2018)

[37] Punit I. S.: Cambridge Analytica's parent firm proposed a massive political machine for India's 2014 elections. In: Reuters (2018)

[38] Punit I. S.: Facebook admits Cambridge Analytica may have accessed the data of over 560,000 users in India. In: Reuters (2018)

[39] Dahir A. L.: 'We'd stage the whole thing': Cambridge Analytica filmed boasting of its role in Kenya's polls. In: Reuters (2018)

[40] Peinado F., Palomo E., Galán J.: The distorted online networks of Mexico's election campaign. In: El País (2018)

[41] Robles R.: How Cambridge Analytica's parent company helped 'man of action' Rodrigo Duterte win the 2016 Philippines election. In: South China Morning Post (2018)

[42] Regulation (EU) 2016/679 of the European Parliament and of the Council on the Protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) https://eur-lex.europa.eu/legal-content/EN/ (2016)

[43] Paul C., Matthews M.: The Russian 'Firehose of Falsehood' Propaganda Model: Why It Might Work and Options to Counter It. In: RAND Corporation (2016)

[44] Kakutani M.: The Firehose of Falsehood: Propaganda and Fake News. In: The Death of Truth: Notes on Falsehood in the Age of Trump (2018)

[45] Harkins S. G., Petty R. E.: The Multiple Source Effect in Persuasion: The Effects of Distraction. In: Personality and Social Psychology Bulletin 7(4) (1981)

[46] Harkins S. G., Petty R. E.: Information Utility and the Multiple Source Effect. In: Journal of Personality and Social Psychology 52(2) (1987)

[47] Flanagin A. J., Metzger M. J.: Trusting Expert- Versus User-Generated Ratings Online: The Role of Information Volume, Valence, and Consumer Characteristics. In: Computers in Human Behavior 29(4) (2013)

[48] Alba J. W., Marmorstein H.:The Effects of Frequency Knowledge on Consumer Decision Making. In: Journal of Consumer Research 14(1) (1987)

[49] Shearer E., Matsa K. E.: News Use Across Social Media Platforms 2018. In: Pew Research Center. https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ (2018)

[50] Shearer E.: Social media outpaces print newspapers in the U.S. as a news source. In: Pew Research Center. https://www.pewresearch.org/social-media-outpaces-print-newspapers/ (2018)

[51] Mitchell A., Jurkowitz M., Oliphant J. B., Shearer E.: Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable. In: Pew Research Center. https://www.journalism.org/americans-who-mainly-get-their-news-on-social-media (2020)

[52] Goertzel T.: Belief in conspiracy theories. In: Political Psychology 15(4), 731–742 (1994)

[53] Brotherton R., French C. C., Pickering A. D.: Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale. In: Frontiers in Psychology (4), 279 (2013)

[54] Brotherton R., French C. C.: Belief in Conspiracy Theories and Susceptibility to the Conjunction Fallacy. In: Applied Cognitive Psychology 28(2), 238–248 (2014)

[55] Thresher-Andrews C.: An introduction into the world of conspiracy. In: PsyPAG Quarterly 88, 5–8 (2013)

[56] Giachanou A., Ghanem B., Rosso P.: Detection of Conspiracy Propagators using Psycho-linguistic Characteristics. In: Journal of Information Science https://doi.org/10.1177/0165551520985486 (2021)

[57] The Making of QAnon: A Crowdsourced Conspiracy. In: Bellingcat - The Q Origins Project. https://www.bellingcat.com/news/americas/2021/01/07/the-making-of-qanon-a-crowdsourced-conspiracy/ (2021)

[58] Alt Right: A Primer about the New White Supremacy. In: Anti-Defamation League. https://www.adl.org/alt-right-the-new-white-supremacy (2017)

[59] Sands G.: What to Know About the Worldwide Hacker Group 'Anonymous'. In: ABC News. https:/abcnews.go.com/hacker-group-anonymous/story?id=37761302 (2016)

[60] Barry D., McIntire M., Rosenberg M.: 'Our President Wants Us Here': The Mob That Stormed the Capitol". In: The New York Times. https://www.nytimes.com/capitol-rioters (2021)

[61] Neilson S., McFall-Johnsen M.: Several groups of extremists stormed the Capitol on Wednesday. Here are some of the most notable individuals, symbols, and groups. In: The Business Insider. https:/www.businessinsider.com/symbols-and-extremist-groups-at-the-us-capitol (2021)

[62] The saga of 'Pizzagate': The fake story that shows how conspiracy theories spread. In: BBC News. https://www.bbc.com/news/blogs-trending-38156985 (2016)

[63] Barkun M.: A Culture of Conspiracy: Apocalyptic Visions in Contemporary America. In: University of California Press, 3–4 (2003)

[64] Douglas K. M., Sutton R. M.: Does it take one to know one? Endorsement of conspiracy theories is influenced by personal willingness to conspire. In: British Journal of Social Psychology 10(3), 544–552 (2011)

[65] Reddit will not ban 'distasteful' content, chief executive says. In: BBC News. https://www.bbc.com/news/technology-19975375 (2012)

[66] Statt. N: Reddit CEO says racism is permitted on the platform, and users are up in arms. In: The Verge. https://www.theverge.com/reddit-ceo-racist-slurs-are-okay (2018)

[67] Fordham E.: Alternative social media Parler added 1M users in a week, still not profitable: CEO. In: Fox Business. https://www.foxbusiness.com/parler-user-numbers (2020)

[68] How fake news is creating profits. In: Adperfect. http://www.adperfect.com/how-fake-news-is-creating-profits/ (2017)

[69] Napley K: The Impact of FakeNews: Politics. https://www.lexology.com/the_impact_of_fake_news_politics (2017)

[70] Exposing Russias Effort to Sow Discord Online: The Internet Research Agency and Advertisements. In: US House of Representatives. https://democrats-intelligence.house.gov/social-media-content/ (2018)

[71] 'Adam Sandler Death Hoax'. In: Snopes. https:/www.snopes.com/adam-sandler-death-hoax (2017)

[72] Higgins E.: Fake news is spiraling out of control - and it is up to all of us to stop it. https://www.ibtimes.co.uk/fake-news-spiralling-out-control (2016)

[73] Kleinman Z.: Fake news 'travels faster', study finds. In: Technology reporter, BBC News. https://www.bbc.com/news/technology-43344256 (2018)

[74] The world's most valuable resource is no longer oil, but data. In: The Economist. https:/www.economist.com/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (2017)

[75] World Economic Forum: Personal Data: The Emergence of a New Asset Class (PDF). (2011)

[76] Grankvist P.: The Big Bubble: How Technology Makes It Harder to Understand the World. In: United Stories Publishing (2018)

[77] Parramore L.: The Filter Bubble. In: The Atlantic. https://www.theatlantic.com/the-filter-bubble (2010)

[78] Bozdag E.: Bias in algorithmic filtering and personalization. In: Ethics and Information Technology 15(3), 209–227 (2013)

[79] Pariser E.: The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin (2011)

[80] Pariser E.: Beware online "filter bubbles" https:/www.ted.com/beware_filter_bubbles (2011)

[81] Gross D.: What the Internet is hiding from you. In: CNN. http:/www.cnn.com/what-the-internet-is-hiding-from-you (2011)

[82] Lai M., Tambuscio M., Patti V., Ruffo G., Rosso P.: Stance Polarity in Political Debates: a Diachronic Perspective of Network Homophily and Conversations on Twitter. In: Data Knowledge Engineering, 124. https://doi.org/10.1016/j.datak.2019.101738 (2019)

[83] Del Vicario M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G., Stanley E. H., Quattrociocchi W.: The spreading of misinformation online. In: Proceedings of the National Academy of Sciences, 113(3):554–559 (2016)

[84] Weisberg J.: Bubble Trouble: Is Web personalization turning us into solipsistic twits?. In: Slate. (2011)

[85] Sunstein C. R.: The law of group polarization. In: Journal of political philosophy 10, 175–195 (2002).

[86] A virtual counter-revolution. In: The Economist. https://www.economist.com/node/16941635 (2010)

[87] Van Alstyne M., Brynjolfsson E.: Electronic Communities: Global Village or Cyberbalkans? (1997)

[88] Van Alstyne M., Brynjolfsson E.: Could the Internet Balkanize Science?. In: Science 274(5292), 1479–1480 (1996)

[89] Del Vicario M., Vivaldo G., Bessi A., Zollo F., Scala A., Caldarelli G., Quattrociocchi W.: Echo chambers: Emotional Contagion and Group Polarization on Facebook. In: Scientific Reports 6 (2016)

[90] Quattrociocchi W., Scala A., Sunstein C. R.: Echo chambers on facebook (2016)

[91] Lee S. T.: Lying to tell the truth: Journalists and the social context of deception. In: Mass Communication Society (2004)

[92] Kaufman P.: Skull in the Ashes. In: University of Iowa Press (2013)

[93] Greenwald G.: Beyond BuzzFeed: The 10 Worst, Most Embarrassing U.S. Media Failures on the Trump-Russia Story. In: The Intercept. https://theintercept.com/beyond-buzzfeed-the-10-worst-most-embarrassing-u-s-media-failures-on-the-trumprussia-story (2019)

[94] Timberg G.: Spreading fake news becomes standard practice for governments across the world. In: The Washington Post. https://www.washingtonpost.com/spreading-fake-news-becomes-standard-practice-for-governments-across-the-world (2017)

[95] Shaw T., Youngblood D. J.: Cinematic Cold War: The American and Soviet Struggle for Hearts and Minds. In: University Press of Kansas (2010)

[96] Lim J.: The language of cross-strait tensions. In: The Interpreter. https://www.lowyinstitute.org/the-interpreter/language-cross-strait-tensions

[97] Luo M.: How the NRA Manipulates Gun Owners and the Media. In: The New Yorker. https://www.newyorker.com/how-nra-manipulates-gun-owners-and-the-media (2017)

[98] Luo M.: N.R.A. Stymies Firearms Research, Scientists Say. In: The New York Times. https://www.nytimes.com/26guns.html (2011)

[99] Awan I.: Cyber-Extremism: Isis and the Power of Social Media. In: Society 54. (2017)

[100] Al-khateeb S., Agarwal N.: Examining botnet behaviors for propaganda dissemination : A case study of isil's beheading videos-based propaganda. In: ICDMW (2015)

[101] Allam H.: Twitter Struggling To Shut Down Bot And Impersonation Accounts Created By ISIS. In: NPR. https://www.npr.org/twitter-to-shut-down-bot-created-by-isis (2019)

[102] Varol O., Ferrara E., Davis C. A., Menczer F., Flammini A.: Online Human-Bot Interactions: Detection, Estimation, and Characterization. arXiv:1703.03107 (2017)

[103] Rangel F., Rosso P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter. In: L. Cappellato, N. Ferro, D. E. Losada and H. Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, vol. 2380 (2019)

[104] Stella M., Ferrara E., De Domenico M.: Bots increase exposure to negative and inflammatory content in online social systems. Proc. of

the National Academy of Sciences of the United States of America, 115(49):12435–12440 (2018)

[105] Boshmaf Y., Muslukhov I., Beznosov K., Ripeanu M.: The socialbot network: when bots socialize for fame and money. In: ACSAC (2011)

[106] Jenna Abrams: the Trump-loving Twitter star who never really existed. In: The Guardian. https://www.theguardian.com/jenna-abrams-the-trump-loving-twitter-star-who-never-existed (2017)

[107] Collins B., Cox J.: Jenna Abrams, Russia's Clown Troll Princess, Duped the Mainstream Media and the World. In: The Daily Beast. https://www.thedailybeast.com/jenna-abrams-russias-clown-troll-princess-duped-the-mainstream-media-and-the-world (2017)

[108] Mihaylov T., Georgiev G., Nakov P.: Finding Opinion Manipulation Trolls. In: News Community Forums. In: CoNLL (2015)

[109] Andjelovic T., Buckels E. E., Paulhus D. L., Trapnell P. D.: Internet trolling and everyday sadism: Parallel effects on pain perception and moral judgment. In: Journal of Personality 87(2), 328–340 (2019)

[110] Navarro-Carrillo G., Torres-Marín J., Carretero-Dios H.: Do trolls just want to have fun? Assessing the role of humor-related traits in online trolling behavior. In: Computers in Human Behavior 114(106551), 1–9 (2021)

[111] Buckels E. E.: Probing the Sadistic Minds of Internet Trolls. In: Society for Personality and Social Psychology (2019)

[112] March E., Steele G.: High Esteem and Hurting Others Online: Trait Sadism Moderates the Relationship Between Self-Esteem and Internet Trolling. In: Cyberpsychology, Behavior, and Social Networking 23(7), 441–446 (2020)

[113] Ghanem B., Buscaldi D., Rosso P.: TexTrolls: Identifying Trolls on Twitter with Textual and Affective Features. In: Proc. Workshop on Online Misinformation and Harm-Aware Recommender Systems (OHARS), Co-located with RecSys 2020, CEUR Workshop Proceedings. CEUR-WS.org 2758, 4-22

[114] Chen C., Wu K., Srinivasan V., Zhang X.: Battling the internet water army: Detection of hidden paid posters. In: ASONAM (2013)

[115] Zannettou S., Caulfield T., De Cristofaro E., Sirivianos M., Stringhini G., Blackburn J.: Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web (2018)

[116] Shu K., Sliva A., Wang S., Tang J., Liu H.: Fake News Detection on Social Media: a Data Mining Perspective. In: ACM SIGKDD Explorations Newsletter 19 (2017)

[117] Ward A., Ross L., Reed E., Turiel E., Brown T.: Naive realism in everyday life: Implications for social conflict and misunderstanding. In: Values and knowledge, 103–135 (1997)

[118] Nickerson R. S.: Confirmation bias: A ubiquitous phenomenon in many guises. In: Review of general psychology 2(2), 175 (1998)

[119] Linfield S.: The Cruel Radiance: Photography and Political Violence. University of Chicago Press (2011)

[120] Sontag S.: On Photography. In: New York Review of Books (1975)

[121] Adatto K.: Picture Perfect: Life in the Age of the Photo Op. Princeton University Press (2008)

[122] Giachanou A., Ríssola E., Ghanem B., Crestani F., Rosso P.: The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers. In: Proceedings 25th International Conference on Applications of Natural Language to Information Systems, NLDB-2020, Springer-Verlag, LNCS(12089), 181-192 (2020)

[123] Giachanou A., Zhang G., Rosso P.: Multimodal Fake News Detection with Textual, Visual and Semantic Information. In: Proceedings 23rd International Conference on Text, Speech and Dialogue, TSD-2020, Springer-Verlag, LNAI(12284), 30-38 (2020)

[124] Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying Misinformation in Microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1589–1599. EMNLP '11 (2011)

[125] Gentzkow M., Shapiro J. M., Stone D. F.: Media bias in the marketplace: Theory. In: Technical report, National Bureau of Economic Research (2014)

[126] Vlachos A., Riedel S.: Fact checking: Task definition and dataset construction.In: ACL'14 (2014)

[127] Yi S. K. M., Steyvers M., Lee M. D., Dry M. J.: The Wisdom of the Crowd in Combinatorial Problems. In: Cognitive Science 36(3), 452–470 (2012)

[128] Mihaylova T., Nakov P., Márquez L., Barrón-Cedeño A., Mohtarami M., Karadzhov G., Glass J.: Fact checking in community forums. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

[129] Popat K., Mukherjee S., Strötgen J., Weikum G.: Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In: Proceedings of the 26th International Conference on World Wide Web Companion, 1003–1012 (2017)

[130] Banko M., Cafarella M. J., Soderland S., Broadhead M., Etzioni O.: Open information extraction from the web. In: IJCAI'07 (2007)

[131] Yates A., Cafarella M., Banko M., Broadhead M., Etzioni O.: TextRunner: Open Information Extraction on the Web. In: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (2007)

[132] Hassan N., Adair B., Hamilton J., Li C., Tremayne M., Yang J., Yu C.: The Quest to Automate Fact-Checking. In: Proceedings of the 2015 Computation+Journalism Symposium (2015)

[133] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2017)

[134] Zlatina M., Teyssou D., Sarris N., Spangenberg J., Papadopoulos S., Alaphilippe A., Bontcheva K.: WeVerify: Wider and Enhanced Verification for You - Project Overview and Tool Demonstration (2019)

[135] Ehrlinger L., Wöß W.: Towards a Definition of Knowledge Graphs. In: Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems - SEMANTiCS2016 and 1st International Workshop on Semantic Change Evolving Semantics (SuCCESS16), 13–16 (2016)

[136] Shi B., Weninger T.: Discriminative Predicate Path Mining for Fact Checking in Knowledge Graphs. In: Knowledge-Based Systems. (2016)

[137] Ciampaglia G. L., Shiralkar P., Rocha L., Bollen J., Menczer F., Flammini A.: Computational fact checking from knowledge networks. arXiv:1501.03471 (2015)

[138] Shiralkar P., Flammini A., Menczer F., Ciampaglia G. L.: Finding streams in knowledge graphs to support fact checking. arXiv:1708.07239 (2017)

[139] Pan J. Z., Pavlova S., Li C., Li N., Li Y., Liu J.: Content Based Fake News Detection Using Knowledge Graphs. In: Proceedings 17th International Semantic Web Conference - Part I (2018)

[140] Bordes A., Usunier N., Garcia-Durán A., Weston J., Yakhnenko O.: Translating Embeddings for Modeling Multi-relational Data. https://proceedings.neurips.cc/paper (2013)

[141] Greco S., Ehrgott M., Figueira J. R.: Multiple criteria decision analysis - State of the Art Surveys. Springer (2016)

[142] Pasi G., De Grandis M., Viviani M.: Decision Making over Multiple Criteria to Assess News Credibility in Microblogging Sites. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) 10.1109/FUZZ48607.2020.9177751 (2020)

[143] Yagerand R.R., Kacprzyk J.: The ordered weighted averaging operators: theory and applications. Springer Science (2012)

[144] Marrara S., Pasi G., Viviani M.: Aggregation operators in Information Retrieval. In: Fuzzy Sets and Systems 324. 10.1016/j.fss.2016.12.018 (2016)

[145] Damiani E., Viviani M.: Trading anonymity for influence in open communities voting schemata. In: 2009 International Conference on Social Informatics (SOCINFO'09),63–67 (2009)

[146] Ceravolo P., Damiani E., Viviani M.: Adding a Peer-to-Peer Trust Layer to Metadata Generators. In: Lecture Notes in Computer Science 3762, 809-815. 10.1007/11575863102 (2005)

[147] Dhar V.: Data science and prediction. In: Communications of the ACM 56(12), 64–73 (2013)

[148] Buntain C., Golbeck J.: Automatically Identifying Fake News in Popular Twitter Threads, 208-215. 10.1109/SmartCloud.2017.40 (2017)

[149] Castillo C., Mendoza M., Poblete B.: Predicting information credibility in time-sensitive social media. In: Internet Research 23(5), 560–588 (2013)

[150] Harris Z.: Distributional Structure. In: Word 10(2/3), 146–62 (1954)

[151] Dunning T.: Statistical Identification of Language. In: Technical Report MCCS, 94–273 (1994)

[152] Afroz S., Brennan M., Greenstadt R.: Detecting hoaxes, frauds, and deception in writng style online. In: ISSP'12 (2012)

[153] Mendoza M., Poblete B., Castillo C.: Twitter Under Crisis: Can We Trust What We RT?. In: Proceedings of the First Workshop on Social Media Analytics, ser. SOMA '10, 71–79 (2010)

[154] Wang W. Y.: "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. arXiv:1705.00648 (2017)

[155] Feng S., Banerjee R., Choi Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (2) (2012)

[156] Rubin V. L., Lukoianova T.: Truth and deception at the rhetorical structure level. In: Journal of the Association for Information Science and Technology 66(5), 905–917 (2015)

[157] NRC Word-Emotion Association Lexicon https://saifmohammad.com/NRC-Emotion-Lexicon.htm

[158] SenticNet Website https://sentic.net

[159] Baccianella S., Esuli A., Sebastiani F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of LREC, 2200–2204 (2010)

[160] Ghanem B., Rosso P., Rangel F.: An Emotional Analysis of False Information in Social Media and News Articles. In: ACM Transactions on Internet Technology (TOIT) 20(2), 1-18

[161] Nakashole N., Mitchell T. M.: Language-Aware Truth Assessment of Fact Candidates. https://www.aclweb.org/P14-1095.pdf (2014)

[162] Boididou C., Andreadou K., Papadopoulos S., Dang-Nguyen D. T., Boato G., Riegler M., Kompatsiaris Y.: Verifying multimedia use at mediaeval 2015. In: MediaEval (2015)

[163] Mahdian B., Saic S.: Using noise inconsistencies for blind image forensics. In: Image and Vision Computing 27(10), 1497–1503 (2009)

[164] Muhammad G., Al-Hammadi M. H., Hussain M., Bebis G.: Image forgery detection using steerable pyramid transform and local binary pattern. In: Machine Vision and Applications 25(4), 985–995 (2014)

[165] Unser M., Chenouard N., Van De Ville D.: Steerable Pyramids and Tight Wavelet Frames. In: Image Processing 20(10),2705-2721 (2011)

[166] He D. C., Wang L.: Texture Unit, Texture Spectrum, And Texture Analysis. In: Geoscience and Remote Sensing, IEEE Transactions (28),509-512 (1990)

[167] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial nets. In: Advances in neural information processing systems, 2672–2680 (2014)

[168] Nataraj L., Mohammed T. M., Manjunath B., Chandrasekaran S., Flenner A., Bappy J. H., Roy-Chowdhury A. K.: Detecting GAN-generated fake images using co-occurrence matrices (2019)

[169] McCloskey S., Albright M.: Detecting GAN-generated imagery using color cues (2018)

[170] Bianchi T., Piva A.: Image forgery localization via block-grained analysis of jpeg artifacts. In: IEEE Transactions on Information Forensics and Security 7(3), 1003– 1017 (2012)

[171] Li W., Yuan Y., Yu N.: Passive detection of doctored jpeg image via block artifact grid extraction. In: Signal Processing 89(9), 1821–1829 (2009)

[172] Qi P., Cao J., Yang T., Guo J., Li J.:Exploiting multi-domain visual information for fake news detection. In: 19th IEEE International Conference on Data Mining (2019)

[173] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image fea- tures for microblogs news verification. IEEE Transactions on Multimedia, 19(3):598–608, 2017.

[174] Jin Z., Cao J., Zhang Y., Zhou J., Tian Q.: Novel visual and statistical image features for microblogs news verification. In: IEEE Transactions on Multimedia 19(3), 598–608 (2017)

[175] Wu K., Yang S., Zhu K. Q.: False rumors detection on Sina Weibo by propagation structures. In: 2015 IEEE 31st International Conference on Data Engineering, 651–662. IEEE (2015)

[176] West J., Ventura D., Warnick S.: Spring Research Presentation: A Theoretical Foundation for Inductive Transfer (2007)

[177] Qi P., Cao J., Yang T., Guo J., Li J.: Exploiting multi-domain visual information for fake news detection. In: 19th IEEE International Conference on Data Mining (2019)

[178] Castillo C., Mendoza M., Poblete B.: Information credibility on twitter. In: WWW'11 (2011)

[179] Tan C., Danescu-Niculescu M. M., Niculae V., Danescu-Niculescu C. M., Lee L.: Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In: Proceedings of the 25th International Conference on World Wide Web, s, 613–624 (2016)

[180] Yang F., Liu Y., Yu H., Yang M.: Automatic detection of rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics,13. (2012)

[181] Ma J., Gao W., Wei Z., Lu Y., Wong K.: Detect rumors using time series of social context information on microblogging websites. In: CIKM'15 (2015)

[182] Ruchansky N., Seo S., Liu Y.: A hybrid deep model for fake news (2017)

[183] Jin Z., Cao J., Zhang Y., Luo J.: News verification by exploiting conflicting social viewpoints in microblogs. In: AAAI'16 (2016)

[184] Ma J., Gao W., Mitra P., Kwon S., Jansen B. J., Wong K., Cha M.: Detecting rumors from microblogs with recurrent neural networks. https://www.ijcai.org/Proceedings/16/Papers/537.pdf

[185] Kwon S., Cha M., Jung K., Chen W., Wang Y.: Prominent features of rumor propagation in online social media. In: ICDM'13, 1103–1108 (2013)

[186] Bessi A., Ferrara E.: Social bots distort the 2016 us presidential election online discussion. In: First Monday 21(11) (2016)

[187] Tacchini E, Ballarin G., Della Vedova M. L., Moret S., De Alfaro L.: Some like it hoax: Automated fake news detection in social networks (2017)

[188] Conroy N., Rubin V. L., Chen Y.: Automatic deception detection: Methods for finding fake news. In: Proceedings of the Association for Information Science and Technology 52(1), 1-4 (2015)

[189] Argamon S., Koppel M., Fine J., Shimoni A. R.: Gender, genre, and writing style in formal written texts. In: Text Talk 23(3), 321–346 (2003)

[190] Koppel M., Argamon S., Shimoni A. R.: Automatically categorizing written texts by author gender. In: Literary and linguistic computing 17(4), 401–412 (2002)

[191] Burger J.D., Henderson J., Kim G., Zarrella G.: Discriminating gender on twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Lan- guage Processing,1301–1309 (2011)

[192] Schwartz H. A., Eichstaedt J. C., Kern M. L., Dziurzynski L., Ramones S. M., Agrawal M., Shah A., Kosinski M., Stillwell D., Seligman M. E.: Personality, gender, and age in the language of social media: The open-vocabulary approach. In: PloS one 8(9) (2013)

[193] Johansson F.: Supervised Classification of Twitter Accounts Based on Textual Content of Tweets. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)

[194] Giglou H. B.,Razmara J., Rahgouy M., Sanaei M.: LSACoNet:A Combination of Lexical and Conceptual Features for Analysis of Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[195] Hashemi A., Zarei M. R., Moosavi M. R., Taheri M.: Fake News Spreader Identification in Twitter using Ensemble Modeling Notebook for PAN at CLEF 2020. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[196] Shashirekha H. L., Balouchzahi F.: ULMFiT for Twitter Fake News Profiling. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[197] Labadie R., Castro D. C., Bueno R. O.: Fusing Stylistic Features with Deep- learning methods for Profiling Fake News Spreader. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[198] Ikade C., Savoy J.: UniNE at PAN-CLEF 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[199] ortennhuemer C., Zangerle E.: A Multi-Aspect Classification Ensemble Approach for Profiling Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[200] Fersini E., Armanini J., D'Intorni M.: Profiling Fake News Spreaders: Stylometry, Personality, Emotions and Embeddings. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[201] Agirrezabal M.: KU-CST at the Profiling Fake News spreaders Shared Task. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[202] Sandoval L. G. M., Puertas E., Quimbaya A. P., Valencia J. A. A.: Assembly of Polarity, Emotion and User Statistics for Detection of Fake Profiles. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[203] Saeed U., Fahim H., Shirazi F.: Profiling Fake News Spreaders on Twitter using Deep Learning LSTM and BI-LSTM Approach. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[204] Bakhteev O., Ogaltsov A., Ostroukhov P.: Fake News Spreader Detection using Neural Tweet Aggregation. In: Cappellato L., Eickhoff C., Ferro N., Neveol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers (2020)

[205] Buda J., Bolonyai F.: An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR- WS.org (2020)

[206] Pizarro J.: Using N-grams to detect Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[207] Walker S. H., Duncan D.B.: Estimation of the probability of an event as a function of several independent variables. In: Biometrika 54(1/2), 167–178 (1967)

[208] Yi D., Ji S., Bu S.: An enhanced optimization scheme based on gradient descent methods for machine learning. In: Symmetry 11(7), 942 (2019)

[209] Weisstein E. W: Norm. From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/Norm.html

[210] Cortes C., Vapnik V. N.: Support-vector networks. In: Machine Learning 20(3), 273–297 (1995)

[211] Breiman L.: Random Forests. In: Machine Learning 45(1), 5–32 (2001)

[212] Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning. Springer. (2008)

[213] Hastie T., Tibshirani R., Friedman J. H.: 10. Boosting and Additive Trees. In: The Elements of Statistical Learning, Springer, 337–384 (2009)

[214] Piryonesi S. M., El-Diraby T. E.: Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling. In: Journal of Infrastructure Systems 27(2) (2021)

[215] Chang C., Lin C.: LIBSVM: A library for support vector machines. In: ACM Transactions on Intelligent Systems and Technology 2(3) (2011)

[216] Fan R., Chang K., Hsieh C., Wang X., Lin C.: LIBLINEAR: A Library for Large Linear Classification (2020)

[217] Giachanou, A., Rosso, P., Crestani, F.: Leveraging Emotional Signals for Credibil- ity Detection. In: Proceedings of the 42nd International

ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 877–880 (2019)

[218] Pennington J., Socher R., Manning C. D.: GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP), 1532–1543 (2014)

[219] Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 (2012)

[220] Nair V., Hinton G. E.: Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th International Conference on Machine Learning (2010)

[221] Diederik K, Ba J.: Adam: A method for stochastic optimization. In: ICLR 2015 (2015)

[222] Vogel I., Meghana M.: Fake News Spreader Detection on Twitter using Character N-Grams. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[223] Koloski B., Pollak S., Skrlj B.: Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[224] Fernández J. L., Ramirez J. A. L.: Approaches to the Profiling Fake News Spreaders on Twitter Task in English and Spanish. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR- WS.org (Sep 2020)

[225] Pinnaparaju N., Indurthi V., Varma V.: Identifying Fake News Spreaders in Social Media. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[226] Lichouri M., Abbas M., Benaziz B.: Profiling Fake News Spreaders on Twitter based on TFIDF Features and Morphological Process Notebook for PAN at CLEF 2020. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[227] Manna R., Pascucci A., Monti J.: Profiling Fake News Spreaders through Stylom- etry and Lexical Features. UniOR NLP @PAN2020. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[228] Niven T., Kao H. Y., Wang H. Y.: Profiling Spreaders of Disinformation on Twit- ter: IKMLab and Softbank Submission. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR- WS.org (Sep 2020)

[229] Russo I.: Sadness and Fear: Classification of Fake News Spreaders Content on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[230] Cardaioli M., Cecconello S., Conti M., Pajola L., Turrin F.:FakeNewsSpreaders Profiling Through Behavioural Analysis. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR- WS.org (Sep 2020)

[231] Chilet L., Martí P.: Profiling Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[232] Majumder S. B., Das D.: Detecting Fake News Spreaders on Twitter Using Uni- versal Sentence Encoder. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[233] Das K.A., Baruah A., Barbhuiya F.A., Dey K.: Ensemble of ELECTRA for Profiling Fake News Spreaders. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[234] Baruah A., Das K., Barbhuiya F., Dey K.: Automatic Detection of Fake News Spreaders Using BERT. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[235] Wu S.H., Chien S.L.: A BERT based Two-stage Fake News Spreader Profiling System. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[236] Espinosa M. S., Centeno R., Rodrigo A.: Analyzing User Profiles for Detection of Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)

[237] Burbach L., Halbach P., Ziefle M., Calero Valdez A.: Who Shares Fake News in Online Social Networks? In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. pp. 234–242. UMAP '19 (2019)

[238] Heinstrom J.: Five Personality Dimensions and Their Influence on Information Behaviour. Information research 9(1), 9–1 (2003)

[239] Ross C., Orr E. S., Sisic M., Arseneault J. M., Simmering M. G., Orr R. R.: Per- sonality and Motivations Associated with Facebook Use. Computers in Human Behavior 25(2), 578–586 (2009)

[240] Rangel F., Giachanou A., Ghanem B., Rosso P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 2696 (2020)

[241] Pennebaker J. W., Boyd R. L., Jordan K., Blackburn K.: The Development and Psychometric Properties of LIWC 2015. Tech. rep. (2015)

[242] Cronbach L. J.: Coefficient alpha and the internal structure of tests. In: Psychometrika, Springer Science and Business Media LLC. 16 (3), 297–334 (1951)

[243] Allen M., Yen W.: Introduction to Measurement Theory (1979)

[244] Roget P. M.: Roget's thesaurus of English words and phrases. Longman (1982)

[245] Watson D., Clark L. A., Tellegen A.: Development and validation of brief measures of positive and negative affect: The PANAS scales. In: Journal of Personality and Social Psychology 54(6), 1063–1070 (1988)

[246] Christopher D. M., Surdeanu M., Bauer J., Finkel J., Bethard S. J., McClosky D.: The Stanford CoreNLP Natural Language Processing Toolkit (2003)

[247] LeFebvre L., LeFebvre L., Blackburn K., Boyd R.: Student Estimates of Public Speaking Competency: The Meaning Extraction Helper and Video Self-evaluation. In: Communication Education (64)3, 261-279 (2015)

[248] John, O.P., Srivastava, S.: The Big-five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In: Handbook of Personality: Theory and Research, pp. 102–138 (1999)

[249] Rothmann S., Coetzer E. P.: The big five personality dimensions and job performance. In: SA Journal of Industrial Psychology (29) (2003)

[250] Costa P. T., McCrae R. R.: NEO personality Inventory professional manual. In: Psychological Assessment Resources (1992)

[251] McCrae R. R., John O. P.: An introduction to the Five-Factor Model and its applications. In: Journal of Personality 60(2), 175–215 (1992)

[252] Ozer D. J., Benet-Martínez V.: Personality and the prediction of consequential outcomes. In: Annual Review of Psychology (57), 401–421 (2006)

[253] Toegel G., Barsoux J. L.: How to become a better leader. In: MIT Sloan Management Review 53(3), 51–60 (2012)

[254] Carter N. L., Guan L., Maples J. L., Williamson R. L., Miller J. D.: The downsides of extreme conscientiousness for psychological wellbeing: The role of obsessive compulsive tendencies. In: Journal of Personality (4), 510–522 (2015)

[255] Matsumoto D., Juang L.: Culture and Psychology: 5th Edition, 271. (2012)

[256] Bamford J. M. S., Davidson J. W.: Trait Empathy associated with Agreeableness and rhythmic entrainment in a spontaneous movement to music task: Preliminary exploratory investigations. In: Musicae Scientiae 23(1), 5–24 (2017)

[257] Song Y.: Associations between empathy and big five personality traits among Chinese undergraduate medical students. In: PLOS ONE 12(2) (2017)

[258] Kaufman S. B., Yaden D. B., Hyde E., Tsukayama E.: The Light vs. Dark Triad of Personality: Contrasting Two Very Different Profiles of Human Nature. In: Frontiers in Psychology (10), 467 (2019)

[259] Friedman H., Schustack M.: Personality: Classic Theories and Modern Research (2016)

[260] Jeronimus B.F., Riese H., Sanderman R., Ormel J.: Mutual reinforcement between neuroticism and life experiences: a five-wave, 16-year study to test reciprocal causation. In: Journal of Personality and Social Psychology 107(4), 751–64 (2014)

[261] Norris C. J., Larsen J. T., Cacioppo J. T.: Neuroticism is associated with larger and more prolonged electrodermal responses to emotionally evocative pictures. In: Psychophysiology 44(5), 823–6 (2007)

[262] Jeronimus B. F., Kotov R., Riese H., Ormel, J.: Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 participants. In: Psychological Medicine 46(14), 2883–2906

[263] Ormel J., Jeronimus B.F., Kotov M., Riese H., Bos E. H., Hankin B.: Neuroticism and common mental disorders: Meaning and utility of a complex relationship. In: Clinical Psychology Review 33(5), 686–697 (2013)

[264] Carducci B. J.: The Psychology of Personality: Viewpoints, Research, and Applications, 173–174 (20)

[265] Neuman, Y., Cohen, Y.: A Vectorial Semantics Approach to Personality Assessment. Scientific Reports 4(1) (2014)

[266] Simonyan K., Zisserman A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 (2015)

[267] He K., Zhang X., Ren S., Sun J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 (2015)

[268] Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.: Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 (2015)

[269] Chollet F.: Xception: Deep Learning with Depthwise Separable Convolutions arXiv:1610.02357 (2017)

[270] Bojanowski P., Grave E., Joulin A., Mikolov T.: Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics 5, 135–146 (2017)

[271] Hochreiter S., Schmidhuber J.: Long short-term memory. In: Neural Computation 9(8), 1735–1780 (1997)

[272] Yan S: Understanding LSTM and Its Diagrams. https://medium.com/understanding-lstm (2018)