LUISS

Dipartimento di Impresa e Management

Corso di laurea magistrale in Data Science and Management

Cattedra Machine Learning

Fake News in LLMs: NLP techniques and genetic optimization for fake news detection

Prof. Giuseppe F. Italiano

RELATORE

Prof. Paolo Spagnoletti

CORRELATORE

Martina Crisafulli

756241

CANDIDATO

Anno Accademico 2022/2023

Index

1. Introduction

The research conducted by “We are social” in collaboration with Hootsuite released in October 2022 shows that 93.4% of all people having access to the Internet use social media (Kemp, 2022). Among the main reasons for using social media, in the top five are “reading news stories” (34.6%), “finding content” (30 %), “seeing what’s being talked about” (28.7%). Clearly, during these last years, the usage of social media has converged towards an informative scope. The intensive use of platforms as a source of information was also determined by an increase in shared content. As a consequence, users become content-viewers but at the same content creators. Most of the contents published on the Internet are not supervised, favoring freedom of expression on whichever topic. However, the problem arises when this information circulating on the Web is false or manipulated. The consequences caused by poor sources of information were first noticed during the 2016 American Presidential election and created even more serious problems during the Covid-19 pandemic.

This comprehensive analysis aims to delve deep into the phenomenon of fake news, scrutinizing its various facets from historical contexts to the current influencers that foster its spread. The research will begin by delineating the taxonomy and definitions that characterize fake news, offering a glimpse into its historical context to better grasp its evolution over the years. Then, the research offers a critical analysis of the factors fostering the spread of fake news, in particular the biases, echo chambers, polarization. The close interplay among all these factors shapes individuals’ beliefs and behavior by unconsciously influencing their perception and understanding of reality. Vosoughi et al. (2018) demonstrate that reaching 1500 readers is six times faster for fake news than for fake news and this is mostly due to novelty. Within this complex web, a multitude of players including social bots and trolls actively participate, exacerbating the dissemination of misinformation through well-orchestrated campaigns and manipulating opinion dynamics. This intricate ecosystem has been further complicated by the advent of advanced technological players such as Generative AI, which have opened up new avenues for the creation and propagation of fake news.

These AI systems, equipped with deep learning capabilities, can mimic human-like writing styles and generate content that is often indistinguishable from those created by humans. This has led to a significant escalation in the volume and sophistication of misinformation campaigns, where falsehoods can be disseminated on a massive scale with unprecedented speed and reach.

Moreover, the bots and algorithms are constantly evolving, learning from the vast amounts of data available online to perfect their strategies and techniques. They are capable of identifying and exploiting human biases, effectively creating and nurturing echo chambers where misinformation can thrive unchallenged. For this reason, the second section of the research intends to shed light on the capabilities and underlying technologies of large language models, including the notable GPT series, and their potential implications in the realm of fake news.

It becomes essential to improve the current strategies to limit the phenomenon. Of course, contemporary solutions are a tradeoff between effectiveness and efficiency. Devising a 100% effective method would imply manually fact-checking every piece of information circulating on the Web, but this would not be either scalable or decisive for this issue. It is also worth mentioning that the latency of this approach could not offer real-time solutions. Here, machine learning comes into play. Platform responsibility against disinformation has strengthened during these last years with the introduction of machine learning models capable of limiting social chaos. Given that the only preventive action platforms can do is to educate users on online behavior, algorithms have played an essential role in the fight against fake news, but no technique has been 100% effective. A substantial part of these approaches has focused on content-based fake news detection, but other approaches have proven very useful like network-based models, leveraging the connections of the user who posts on social media and its "friends", or multi-modal models capable of integrating images and videos to textual analysis.

However, to devise a powerful detection system data must be available and consistent, which is usually a very challenging state to achieve. Data quality concerns are indeed one of the main limitations of current approaches (Capuano et al., 2023). Utilization of different datasets can help in having lots of data to increase the model's performance but at the same time, it makes the process of data collection, verification, and storage complicated. Furthermore, most of the time accessibility to information is limited and it is not possible to enrich content-based information with user one. On the other hand, a small dataset may entail other challenges as well, like performance worsening and limited generalization.

Due to the limitations cited above, the final goal of this research is not to devise a more performative model for fake news but providing new insights that can help future researchers in identifying and tackling the multifaceted challenges associated with misinformation spread in a consistently. By comprehensively examining the various factors influencing the generation and propagation of fake news, this research aims to offer a more nuanced understanding of the dynamic landscape of misinformation. Data used in the study was gathered via web scraping from Politifact.com, an independent American fact-checking organization and this was the only source used to guarantee consistency in the dataset and avoiding outliers-related problems. In order to address the imbalance in the dataset created, undersampling was performed on the majority class. Due to the data availability limitation, a news-content based approach was chosen. Consequently, feature selection centered around linguistic and stylistic elements that are intrinsic to the textual content found in news articles. This method leverages the nuances in language usage, such as semantic coherence, syntactic patterns, and stylistic tendencies that might differentiate genuine news from fabricated stories.

Several algorithms were implemented and optimized through the use of genetic algorithm, ranging from traditional models to more complex ones like neural networks and natural language processing techniques (BERT, DistillBERT). The results underscore the pronounced advantage of neural networks and NLP algorithms over traditional methods in deciphering text patterns; a 70% score in overall accuracy performance (for BERT and DistilBert), echoed across other metrics, attests to the model's commendable efficacy even when working with a constrained dataset. Furthermore, the news-content features provide a substantial increase in model performance. The evaluation of these algorithms was not limited to accuracy but focused more on more insightful metrics like ROC curve and recall minimizing overfitting patterns that are often overlooked by research. It is important to notice that on this small dataset, the effect of the scaling laws is evident. In language modeling, scaling laws are an actual optimization method. Indeed, transformer-based models only produce amazing outcomes when the model size and parameter count are large. Using a smaller D, performance stops rising as the model becomes more sophisticated and overfits (Kaplan, 2020).

This assumption holds also for the research by Okunoye et al. (2022) which showed the pivotal role of genetic search applied to deep learning algorithms. Indeed, the research shows an accuracy of 74% on fake news and 56.56% on real news with the test model loss being higher than train loss but avoiding overfitting. The limitations of neural networks underlined by Okunoye et al. (2022) are indeed their large memory requirements and sensitivity to various random weight initializations, in particular for LSTM which make the model prone to overfitting. Considering these considerations, the results obtained by this research on neural networks are satisfying, showing that also on smaller datasets the overfitting issue can be limited achieving a moderate accuracy on fake news (in our case simple LSTM obtained 65% accuracy, 63% on precision, 75% on recall). In conclusion, model complexity and efficacy must be correctly balanced for a good fake news detection system.

Furthermore, the study intends to open up opportunities to expand the dataset, possibly incorporating multilingual and multicultural dimensions, to foster a more inclusive and robust fake news real-time detection system.

This introduction serves to provide readers with a glimpse of the comprehensive analysis they are about to undertake, setting the stage for the detailed explorations in the subsequent sections.

2. FAKE NEWS: UNDERSTANDING THE CONTEXT

This section introduces the fake news environment and how it became headline news in the past decade. First, it is necessary to explain the taxonomy that is employed in the research to clarify what each word refers to. Then, there is an exploration of the major events where fake news played a key role in highlighting the impact of this phenomenon on society.

The broadest part of the chapter deals with the analysis of how fake news is fabricated and how it spreads within society, focusing on its power and velocity.

2.1 Taxonomy and definitions

There does not exist a single definition for fake news. One of the most used definitions in literature has been given by Lazer et al. (2018): "fabricated information that mimics news media content" (Lazer, 2018). However, according to Wang et al. (2019), this might be belittling, indeed this definition seems to shadow satire or politically centered news spreading. The case has even been addressed by the UK Parliament which seems to be in favor of more technical words to describe the phenomenon (i.e., misinformation or disinformation) because the boundaries of this definition are not clear yet and may cause ambiguity (Wang et al., 2019).

According to Wang et al. (2019), source credibility is challenged in current years where user-generated content has exponentially increased, limiting fact-checking solutions. When reading news, three elements come into play: who reads, who interprets and the message.

Rubin et al. (2016) identify three types of fakes: serious fabrications, large-scale hoaxes and humorous fakes. The first type of news is purposedly carried out to highlight specific elements of the news and deceive readers' opinions. For instance, tabloids and the press usually display the so known click baits, headlines created to draw attention to the news without considering the veracity of the information. Hoaxes replicate real existing news, masquerading misleading, and false information as verified ones. Finally, the last type is called humorous because the purpose is to entertain, just like satire.

The concept is usually connected to three important elements: "misinformation", "disinformation" and "malinformation" which form together the information disorder concept (Ruffo G. et al., 2022).

Misinformation denotes misleading information created unintentionally like the click baits while the second identifies content created with the aim of deceiving people (Lazer et al., 2018); as for malinformation, it refers to pieces of information partially or totally true. The research carried out by Ruffo G. et al. (2022) shows that political science, computer science, and psychology are the most related disciplines related to the fake news problem, indeed politics-based news has a high probability of containing wrongful information. Right after the US presidential election, Wardle (2017) published a taxonomy for fake news having 7 types of mis- and disinformation:

1. Satire: entertaining purpose, no harm-oriented

2. Misleading Content: deceptive use of information
3. Imposter Content: when news is impersonated
4. Fabricated Content: new content created to frame individuals
5. False Connection: the headlines written to draw attention do not support the thesis of the text
6. False Context: invent a false context on true news
7. Manipulated Content: intentionally modified piece of information to support a claim

The world of fake news is usually connected to another concept: hoax. Hoaxes are false and outdated pieces of information spontaneously spread by users across the Web (Rahmat and Areni, 2019) both by bots and internet users.

Tambuscio et al. (2015) analysis of misinformation studies how hoaxes work. Hoaxes recall the concept of viruses where every individual is represented by a node that can be "infected" by false news when coming across infected players or "recover" with the help of fact-checking.

The concept of hoaxes became particularly important with the arrival of the 2016 US presidential campaigns which marked the start of the disinformation era, showing the power of "computational propaganda" (Wiesenberg, R. & Tench, R., 2020). Indeed, research on the 2016 elections revealed a massive use of automated accounts to manipulate public discourse and create or destroy beliefs.

The phenomenon of social bots in politics was first studied in 2012/2013 for the US election campaign and then was deepened in the later years leading to the disastrous usage of computational propaganda in the 2016 elections as opinion manipulation instruments (Wiesenberg, R. & Tench, R., 2020). But what is a social bot?

A bot is an automated algorithm that simulates human actions, a well-spread example is the automatic email response. The adjective "social" has been added later to connect this computer program to the world of social media (Wiesenberg, R. & Tench, R., 2020). A bot can be used to perform any of the "online" activities we do every day on our laptop, from automatic emails to chatbots, but it is not necessarily made with a beneficial intent, most of the time the word bot has a negative connotation. However, as the paper specifies "social bots are neither unethical nor ethical", it is how they behave which defines their nature (Wiesenberg, R. & Tench, R., 2020, pag.4). Indeed, the author highlights how their features can be implemented for business strategic purposes; according to 2016 studies, 23.6% of organizations use automation for content distribution and the remaining part for content creation and adaptation (Wiesenberg & Tench, 2020).

McKinsey's report in 2022 testifies that in the cited year the usage of AI adoption in at least one function of organizations has grown by 50% in particular for robotic process automation and computer vision (Chui et al., 2022). Furthermore, in 2023, AI shares have grown so much after the release of ChatGPT and generative AI models that some investors even foresee a tech bubble in the upcoming years. Most importantly, no risk assessment has been pursued yet and the consequences of AI automation keep rising unrestricted by authorities. OpenAI, the company that invented ChatGPT, an AI chatbot, closed at a valuation between $27 and $29 billion in 2023 despite its revenues being much lower (Financial Times, 2023).

With the rise of online-generated content, some of the damaging consequences of fake news have already been experienced. Phenomena like the formation of "echo chambers" and "filter bubbles" where like-minded individuals are shielded from opposing perspectives have been the cause of social chaos in recent years, especially during the Covid-19 pandemic.

With the release of AI-generated content which is just trained on thousands of data and Large Language models (LLMs), the possibility of generating made-up news is much higher. Although they have a lot of potential in the fight against false information, LLMs are a two-edged sword in this conflict. Indeed, they provide useful tools for identifying and battling false information, but it's crucial to be aware that they can also be misused to produce false information on their own. The real objective of this upcoming decade will be maintaining the equilibrium between the exploitation of LLMs potential and risks.

2.2 A bit of historical context

The phenomenon of fake news is not new, it was born in the 19th century. The "Great Moon Hoax" of 1835, when the New York Sun published several articles about the discovery of life on the moon, serves as one historical illustration (Allcott & Gentzkow, 2017). A more recent episode is the "Flemish Secession Hoax" of 2006, in which a public television station in Belgium falsely claimed that the Flemish parliament had declared independence from Belgium. A lot of viewers took this report to be true. Supermarket tabloids like the Weekly World News and the National Enquirer have long been known to publish a mixture of stories that are both partially true and outright false (Allcott & Gentzkow, 2017).

Even though the phenomenon has existed for so long, the greatest consequences of disinformation came in 2016 after the 2016 US presidential elections. These elections represent the apotheosis of political fake news propaganda thanks to the large social media usage by political figures. The fake news scandal known as Pizzagate, which swept social media during the 2016 US Presidential elections, stands out as one of the most memorable examples. In this scandal, it was falsely asserted that Hillary Clinton was the head of a child trafficking ring with a base of operations in a pizzeria. (Ansar, W. & Goswami, S., 2021). The amplified effect of this news degenerated into a man walking inside the pizza restaurant firing bullets.

According to recent research, the COVID-19 pandemic fake news has caused the most harm to people's health in recent months (Pennycook et al., 2020). With the arrival of Covid-19, the psychological distress from the lockdown and the fear of the unknown created a social media environment in which fake news spread faster than ever: starting from social chaos and arriving to violence like the burning of 80 mobile towers in response to the 5G conspiracy theories (Gupta, A. et al., 2022). When Moscadelli et al. (2020) examined the news about COVID-19 in Italy, they discovered that fake news had been reported more than 2 million times—more than 23% of all the articles in their study. The effect of Covid-19 on society and information was so devastating that it fully earned the nickname of “infodemic” because of the enormous amount of information flooding, especially the unreliable ones. Most of the platforms tried to mitigate the consequences of poor sources of information by trying to label deceiving news, assuming that warnings might convince people to stop reading fake articles (Gupta et al., 2022). The inoculation theory was put forth by van der Linden et al. (2020) in the context of avoiding persuasion. The authors suggested that, similarly to how a virus can be rendered ineffective through immunization, people can be psychologically immunized against false information about COVID-19 by making them aware that such information is being disseminated with bad intentions and by simultaneously exposing them to information debunking false claims. According to Allcot & Gentzkow (2017), even though fake news has a long history, there are several reasons to believe that it is becoming more and more important. First off, entry barriers to the media industry have dramatically decreased as a result of how simple it is now to set up websites and how simple it is to monetize web content through advertising platforms (Allcot & Gentzkow, 2017). Higher entry barriers prevent false reporting because mass media outlets are deterred from doing so due to reputational issues. Second, social media have become increasingly popular. In 2016, active Facebook users per month reached 1.8 billion, and Twitter users approached 400 million (Allcot & Gentzkow, 2017). Third, as the paper reports via a Gallup study, trust and confidence in the American media (just like in the rest of the world) "when it comes to reporting the news fully, accurately, and fairly" have continued to decline. Fake news gaining more traction could be both a cause of and a result of the declining trust in traditional media. Fourth, the growing hostility between each side of the political spectrum- political polarization- and deeply related to echo chamber effects only enhances the spread of manipulated information (Allcot & Gentzkow, 2017).

This is because individuals tend to seek out and consume information that confirms their pre-existing beliefs, rather than challenging them (so-called cognitive bias that will be later described in depth). As a result, they are more likely to share and spread misinformation that aligns with their worldview, contributing to the erosion of trust in traditional media.

However, the consequences of fake news are not always so visible and require a larger consideration of factors in order to be managed. According to Scheibenzuber et al. (2023), the strategic framing of news is essential for mind-shaping purposes. The concept of framing is identified as the action of selecting aspects of a “perceived reality” and highlighting them when communicating a text/opinion or thought with the specific purpose of either promoting them or the opposite (Entman, 1993). Strategic framing has the specific goal of persuading users to believe some fact or opinion; it can be distinguished into (Scheibenzuber, 2023):

- Emotional: tries to trigger emotional reactions
- Value: highlights norms and values to influence people’s behavior
- Semantic: tries to evoke connections and associations regarding a certain topic

According to Scheibenzuber et al. (2023), in the fake news world, the first two are the most commonly used strategies. In order to understand how strategic framing is carried out, it is essential to understand some correlated topics that can help us build a complete understanding of the fake news environment.

2.3 Factors influencing the spread of fake news

2.3.1 Biases

The reason why we are so attracted to fake news is usually independent of the news itself. Indeed, our biases act as filters whenever we read, look, or think of something. Biases can be distinguished into cognitive and social: the former refer to the psychology of the individual while the latter to the ones of the society.

The bandwagon effect, one of the most common social biases, consists of aligning opinions with the ones of the society (Ruffo et al., 2022) and it is based on the social influence theory whose underlying hypothesis is that individuals tend to conform their behaviors to the social group they belong to. The problem is that it is very difficult to realize the non-objectivity (third-person effect) of a social group especially when subject to naive realism- the tendency for people to believe that the reality they perceive is objective and factual (Ruffo et al., 2022). The bandwagon effect is very similar to the homophily theory which states that in networks people sharing similarities have the tendency to stay together facilitating homogeneity and congruence (Ruffo et al., 2022) which lay down the basis for echo chamber effects. From this perspective, unfavorable news from adversarial news sources can be categorized as an evil attempt to undermine a widely held belief. The homophily principle is explained by two lines of reasoning:

1. the theory of self-categorization (Turner, 1987, as cited in Sun et al., 2022)
2. the similarity-attraction hypothesis (Byrne, 1971, as cited in Sun et al., 2022)

According to the similarity-attraction hypothesis, people are drawn to others who are like them because they can reduce psychological discomfort brought on by cognitive or emotional dissonance and offer mutual confirmation of each other's attitudes and beliefs (Byrne, 1971, as cited in Sun et al., 2022). Moreover, according to the theory of self-categorization, people categorize themselves and others according to specific social categories like gender (Turner, 1987, as cited in Sun et al., 2022). People's assessments of how much they and others resemble one another are influenced by this self-categorization process, which in turn affects how they interact with one another.

As for cognitive bias, one person's beliefs can be reinforced by a variety of biases, creating a feedback loop that can further confirm their false perception (Ruffo et al., 2022). According to Ruffo et al. (2022), attentional bias is a mechanism that explains why people are more likely to notice something if they are already thinking about it. If we persuade ourselves of a startling or unsettling fact, we might begin to notice it more frequently. One of the most famous is the confirmation bias: it works as a reinforcing filter which means that we read attentively only what confirms their existing beliefs not what opposes them (Ruffo et al., 2022). As a consequence, the uncontrolled bias causes a larger acceptance of information that pleases the reader, also known as desirability bias (Lazer et al., 2018). Ruffo et al. (2022) confirm that people rarely seek disagreement (congruence bias), indeed readers manage to accept only what is in agreement with their preexisting beliefs.

The necessary trait of fake news is the emotional charge: usually the emotional power of the conveyed information, the higher the influence that it has on readers' opinions. Indeed, emotional bias may lead individuals to ignore or accept facts that affect their sentiments positively or negatively: either way the perception of the argument will be distorted (Ruffo et al. 2022). For instance, the research of Ferrara & Yang (2015) studies how negative sentiment in tweets can amplify the oversharing of negative messages by readers. Another bias worth mentioning in Ruffo et al. (2022) is the overconfidence effect which affects people's tendency to overestimate their knowledge of certain topics and to perceive their partial information as highly accurate. This usually results in a backfire effect since being overly confident leaves no space for divergent opinions, which means that challenging or correcting one’s opinion may lead to reinforcing the opposite point of view.

Furthermore, contrasting cognitive biases with fact-checking has not proven to influence positively disinformation.

2.3.2 Echo chamber and polarization

As previously anticipated, cognitive biases unintentionally shape opinions and are at the basis of group polarization. According to Sunstein’s theory (2002), polarization occurs when a member of a group's initial propensity towards an opinion is strengthened by group discussion. The term group polarization defines the development of a group discussion on a particular topic or problem (Sunstein, 2002). Sunstein emphasizes the distinction between cascade effect and polarization. The two are clearly connected but the cascade effect is the process of leading an individual of the group’s opinion towards an established pattern whereas polarization exaggerates this tendency and gives rise to extreme ways of thinking.

While conformity may seem like a significant factor, according to Sunstein (2002), it can be connected, but not essential, to the surfacing of polarization. Instead, one of the most important triggers for polarization can be redirected to social comparison defined as the individuals’ desire to be perceived favorably by their group members. Users adjust their positions in alignment with the prevailing viewpoint. However, upon hearing the perspectives of their peers, they often discover a divergence from their initial position and subsequently shift their own stance accordingly (Sunstein, 2002). As a result, the group's position is pushed toward an extreme, and individual members experience shifts in their own beliefs.

The second factor cited by Sunstein (2002) is persuasiveness, the most convincing argument agreed upon by society wins everybody’s opinion. Of course, not every group discussion dynamics ends up with polarized extremes, and many other factors may come into play.

For instance, information-spreading dynamics is an interplay between personal biases, social biases, and algorithmic filters. Algorithmic filters are often accused of contributing to filter bubbles, restricting users' exposure to diverse opinions (Ruffo et al., 2023).

Indeed, recommendation systems have the goal of maximizing the time spent on a social network/ website and revenues. How do they work? They customize advertising based on users‘online behavior and filter content shown to the user. facilitate the development of virtual environments composed of like-minded people (“echo chambers”), where the user’s opinion is reinforced by content published by users with the same thoughts. According to Ruffo et al. (2023), recommendation systems reflect a particular “propaganda technique" which is repetition: the more a targeted message appears, the higher the reinforcing effect. This greatly reduces the diversity in contents shown. Even though researchers are still uncertain about the direct connection between recommendation algorithms and echo chambers, Geschke et al. (2018) paper proposes a triple-filter bubble framework that explains the emergence of echo chambers by focusing on three filter levels: individual, societal, and technology. Surprisingly, the paper's results reveal that societal and technological levels are equivalent: when social posting is enacted, echo chambers emerge, and the same occurs when recommendation algorithms are put in place. In Figure 1, the graph shows how the relationships across clusters vary; when no social posting is enacted "info-bridges" are heavy whereas when social posting is active these bridges disappear. However, there are doubts about the actual magnitude of algorithms' influence on echo chambers (Ruffo et al., 2022) due to the inability to measure the counterfactual. The research of Möller, et al. (2018) proves that recommendation systems do not lower topic diversity in case users are similar to each other while the opposite occurs in case no past data on the user is provided.

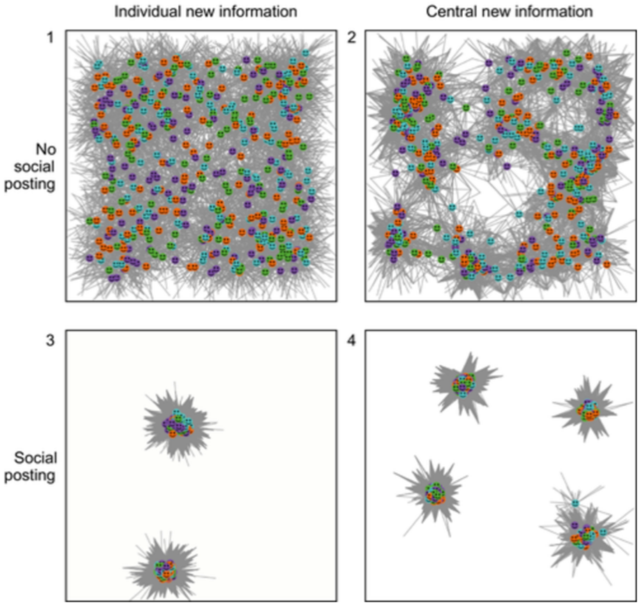


Figure 1- Geschke et al. (2019)

Del Vicario et al. (2016) paper focuses on the behavior of different echo chamber communities and reports that two different chambers can behave in the same way when there is similarity across users' behavior and polarization is dependent on the level of involvement in the community. Indeed, the more active users are the ones driving faster negativity inside communities.

Thus, echo chambers emerge when the level of polarization is at the maximum (Ruffo et al., 2022).

There are multiple examples of polarized groups, for instance, the fight between pro-vax and anti-vax during Covid-19 pandemic or the political rivalry in the famous 2016 US elections which clearly show the chaos and violence caused by these extremes.

2.4 The diffusion of fake news: opinion dynamics

How fast does information travel? According to Ruffo et al., opinions are continuous variables. Supposing x is the opinion and i is an individual, his/her opinion at time t is $x_i(t)$. Information traveling can reach different equilibriums (Ruffo et al., 2022):

1. Consensus: every node has the same opinion at time $t *$
2. Polarization: nodes have opposing opinions
3. Fragmentation: nodes are divided among more than two opinions

Each opinion can be changed according to predetermined rules influenced by polarization and biases. Information spreading models, according to Ruffo et al. (2022), imply also that every node evaluates its neighbors' opinions when close enough but the researcher points out that this is an erroneous belief since the individual is subject to most of the biases mentioned previously. Thus, the nodes, or better the individuals, absorb only part of the information they read due to attentional biases and profiling algorithms.

The traditional opinion dynamics models include the Bounded Confidence Model (BCM) where opinions can be defined as concordant and discordant within a distribution that goes from -1 to 1. The main idea behind the model is that individuals have a confidence interval that limits the range of opinions they are willing to consider. Therefore, individuals can update and consider others' people viewpoints only when within their bounded confidence interval, which means their way of thinking is sufficiently similar (Hegselmann, R., Krause, U., & Riehl, J., 2002).

Quattrociocchi et al. (2014) consider two networks: the gossiper and the media one where the former follows the BCM. Indeed, the opinion of individual i is influenced by a great number of factors, the interaction with individual j, convergence factor μ_{gg}, σ_{gg} which defines the threshold after which gossipers do not communicate and θ the Heaviside's theta function.

The model's results described by Quattrociocchi, W., Caldarelli, G., & Scala, A. (2014) consider two important variables: localization and tolerance. The first describes the equilibrium state of the network, meaning consensus, polarization, or fragmentation while the latter the distance between personal opinion and the received information. The lower this distance the more easily you can get influenced by the news because of your similarity of ideas. Their research shows homogenization of ideas (consensus) in a normal network with unpolarized media is usually reached with high levels of tolerance and the consequent decrease in opinion distance (Quattrociocchi, W., Caldarelli, G., & Scala, A., 2014). In this case, the transition toward consensus is smoothened. In the opposite case, polarized media bring about fragmented opinions. As can be observed by the experiment conducted (Figure 2) the effects are opposite depending on the media competition. In addition, even by considering the information credibility as a variable in the model, the full consensus in the case of polarized media is not reached.

In Ruffo et al. (2022), the model proposed takes into consideration the friending/unfriending mechanisms in social media. The statistical distribution of the opinion of user i is composed by two variables μ, the influence

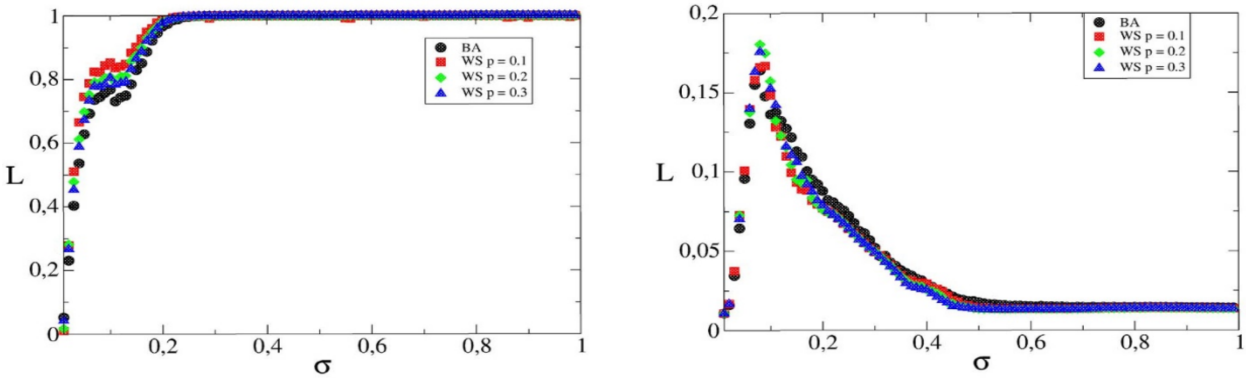


Figure 2 - Quattrociocchi et al. (2014)

strength parameter equal to $1 - g$, the resistance factor and the Kronecker delta δ_{ij} which instead checks for concordance of the opinions between user i and j. The higher the concordance, the more likely it is.

to establish a connection with the new user. As this process continues, the cluster intensifies its density and thus can hinder the diffusion and radicalization of new ideas (Ruffo et al., 2022). Indeed, when fake news enters a very dense cluster the backfire effect could lead to increased polarization.

Furthermore, according to Schelling (1971), even if an agent allows one-third of its neighbors to hold different opinions, the system still tends towards a homogeneous neighborhood.

Now that opinion dynamics have been laid out, it is essential to understand how information travels and why fake news seems to spread faster than every other piece of information.

Moreno et al. (2004) model defines a rumor propagation model composed of three states:

- Ignorant, individuals who have not yet heard the rumor and are therefore susceptible to being informed
- Stifler, individuals who are aware of the rumor but have stopped spreading it
- Spreader, individuals who are actively spreading the rumor

The spreading process starts from the spreaders: when a spreader encounters an ignorant, the ignorant becomes a new spreader with probability λ. The decline of the spreading process can be attributed to either a mechanism of "forgetting" or spreaders realizing that the rumor has lost its "news value."

Consequently, when spreaders interact with other spreaders or stiflers, they become stiflers themselves with a probability of α. The ultimate goal of this model is to maximize the fraction of the population that ultimately learns the rumor. Therefore, we assume that contacts between spreaders are directed, meaning that only the individual initiating the contact loses interest in further propagating the rumor. As a result, there is no double transition to the stifler class. But how does falsehood win over truth?

Vosoughi et al. (2018) analyzed the spread of fake news on Twitter following the development of "cascades". Every time a user posts something, the propagation of the statement through retweets and comments creates a rumor cascade. The research by Vosoughi et al. (2018) characterizes fake rumors according to 4 parameters:

1. Depth, number of retweets
2. Size is the number of users involved in a cascade
3. Maximum Breadth: number of users at each level of depth
4. Structural Virality: as cited by Goel et al. (2015) in Vosoughi (2018) is a measure keeping track of the mean distance between every pair of nodes in a cascade

As expected, the results of the paper show that fake news travels six times faster than truth to reach 1500 readers. A very interesting result is also the depth of a true cascade which never exceeded a depth of 10 contrary to the fake one whose depth attained 19 ten times faster than the former Vosoughi et al.(2018). According to the same paper, truth rarely "infects" more than 1000 people, and the 1% of fake news is capable of reaching between 1000 and 100000 people.

As anticipated by Moreno et al. (2004) research, the decline of a diffusion process can be linked to the news losing its value. Basically, newness is essential in the sharing process: the higher the novelty, the higher the value people attribute to it either because it signals access to unique and rare information or because it helps in decision-making or both.

This is why the reasons for this significant difference can be found in the tweet content itself: contrary to what someone can assert, it is not the number of followers that counts but how alluring the post is (Vosoughi S., Deb Roy, D. & Sinan, A.S., 2018). Indeed, it is the novelty that does the job and increases the shares of information together with bots' power. As a consequence, it is not so surprising that fake rumors generate higher surprise and disgust whereas truth inspires more sadness and trust.

As a result, people are more willing to share falsehood because of its higher "attractiveness".

By studying the spread of information on Facebook, also Ceylan et al. (2023), attribute the spread of fake news to multiple reasons. Recalling Vosough et al. (2018), the first one is the incapacity of verifying a certain claim which drives people to be interested in "novel" headlines and unexpected articles. Then the second reason cited in the paper is bias. Indeed, users are mostly attracted by what confirms their opinion and this is how an echo chamber originates: from similar users. Ceylan et al. (2023) idenfy a third reason to take into account in the spread of misleading content: habits. In the research conducted, once users took the habit of sharing content, also the percentage of false content increased.

Some of the most used tools to combat misinformation were banners and pop-ups signaling the low quality of the content displayed. Ceylan et al. (2023) reveal that, even after knowing the news accuracy, the sharing trend regarding false news remains high. Instead "low-posting" users were found to be more attentive at distinguishing true from false.

But are people so careless about what they share?

Talwar et al. (2019) suggest that the matter is related to trust in social media. The higher the trust, the lower the willingness to check for data authenticity. Unexpectedly, the positive association between fake news and online trust was proven correct: people who rely more on online contents are more willing to take risks when sharing news. Vosoughi et al. (2018) emphasize the role of humans in this diffusion process, due to the fact that the presence of bots in the last decade created some concerns over their role in spreading misinformation.

2.4.1 Social bots, trolls, and humans: pattern of diffusion among online players

According to Vosoughi et al. (2018), the role bots play in spreading misinformation has the same relevance as the human one. This means that the difference in the increase of fake news does not vary substantially when automation comes into play.

First of all: how can we define bots?

The great advances made in technology in this last decade have intensified the presence of algorithms in our lives. Indeed, the simplest way to describe bots is automated accounts driven by algorithms. The misuse of this tool in these last years, especially during the 2016 US presidential elections, has given over the years the word "bot" a negative connotation. As anticipated, this led to the "computational propaganda" era in 2016 and the “infodemic” era during Covid-19. Both are examples of how fake news destroyed society, bringing polarization dynamics to the extremes. Cresci et al. (2020) analysis of bot history highlights how their technical nature has evolved so much over the years making them very difficult to spot. According to Cresci (2020), the first work that addresses bot detection goes back to 2010 and had two particular characteristics, the first one was supervised learning, the second was that accounts were tested one at a time.

Of course, those simplistic models had different limitations on a world that was rapidly growing. For this reason, bots have grown so much in complexity and similarity to humans that right now the line separating them is very blurred.

Luceri et al. (2021) research on 2018 Twitter data identifies two types of accounts hyperactive and ordinary ones based on two parameters: active days and frequency. The threshold that distinguishes the two types of accounts is having both parameters bigger or equal than 1 and they are usually connected to misinformation spreading. What was discovered in the paper is that even when ordinary accounts outperform hyperactive ones in terms of size (only 9% of accounts were hyperactive), they are still capable of covering a larger area of social media (approximately 70% in the paper). 38% of the hyperactive accounts are bots but their activity is so much higher than their size (approximately a bot published three times more than a human user). The fact that the number of active bots never increased substantially and kept a steady pace proves the increasing similarity between bots and human behavior (Luceri et al., 2021).

Differently, the number of bot accounts created suffered a large increase right before the elections and according to Luceri et al. (2021) contributed to 62% of all the conspiracy theories’ narratives.

The diffusion of fake news through online players has also been conducted by Mazza et al. (2019) which identifies three relevant players in the diffusion process: social bots (previously described), and trolls (meaning individuals voluntarily causing chaos by spreading fake news). The study was conducted on a dataset gathered from Twitter during the 2016 US presidential election and consisted of the application of several algorithms capable of distinguishing patterns in the spread of information. The study suggests that trolls and social bots have a significant impact on shaping minds on Twitter and their behavior is largely different from human users. Furthermore, these two players were found to be more willing to post divisive and controversial content than humans and trolls adopt a more aggressive behavior while social bots tend to amplify and disseminate. Indeed, the SHAP analysis carried out by Mazza et al. (2019) reveals some interesting comparisons in the diffusion pattern of information:

1. Trolls vs. humans: troll accounts tend to use a higher number of characters and use more novel words while humans’ posts get replied to more easily.
2. Trolls vs. social bots: the comparison reveals that trolls get more retweets and contain more URLs which implies that the lexicon is carefully chosen aiming at language novelty. However, even if bots are automated accounts, their reply ratio is higher which results in higher visibility on social media.
3. Social bots vs. humans: social bots generate tweets from multiple sources, thus requiring a higher number of characters while humans have limited sources from which to draw.

The results shown in the previous analysis seem to indicate that automated accounts seem to reach more people than another type of accounts.

Do all of the automated accounts behave unethically? The answer is no, but as AI continues to advance, tech stocks rise, and the so-far goal of human-level AI seems approaching. Now that automation is accessing every subset of our everyday life, we must learn to make use of it in the best way.

In addition, bots are incredibly capable of gaining users' trust and central positions in the network.

The issue of people's inability to distinguish between reliable sources and the spread of false information was also highlighted by the unsettling fact that one in three human retweets included sharing content created by bots (Luceri et al. 2021). These results were also confirmed by Erokhin & Komendantova (2023) who focused on the spread of conspiracy theories by bots. The study of conspiracy theories on earthquakes addresses some of the questions already seen in Luceri et al. (2021). Even if most of the number of accounts belonged to humans, bots accounts were the most active in the social network.

Another study by Shao et al. (2018) studies the distribution of low credibility sources. Surprisingly, the virality of the low-credibility facts was as much developed as the fact-checking articles. This occurred because as users kept interacting with a misleading fact, it was noticed that a small number of accounts, known as "super-spreaders," began to control the larger portion of the distribution. According to Shao et al. (2018), the practice of mentioning well-known people in tweets that contain links to unreliable sources is another tactic that bots frequently use. In particular, they also act at the beginning of the spreading of a post increasing right from the start the virality of news (Shao et al., 2018).

According to the research, the employment of bots would certainly diminish the level of low-credibility content and that could help in lowering misinformation across online platforms. However, academic literature does not consider the potential benefits one could draw from the implementation of automatic accounts.

Bots acquired a very negative connotation over the years and their beneficial power has always been overlooked. For instance, nowadays almost every organization employs virtual assistants to help users in their experience on a specific web page: those are nonetheless chatbots. ChatGPT hit one million users in just 5 days and that is also a bot.

So, would prohibiting bots lower fake news? Yes.

Would it be worth eliminating this technology from our lives and stopping its evolution? Considering the big advantages it provides, both from a commercial and non-commercial point of view, the answer here is much more complex. Plus, is it really possible to stop technology?

3. HUMAN-GENERATED AND MACHINE-GENERATED TEXT: GENERATIVE AI

This chapter sets the stage to explore the intricate world of generative AI, delving deep into its fundamentals, its remarkable capabilities, and its influence on various sectors of society. Generative AI, equipped with the ability to produce content that mirrors human output, harbors the potential to be a double-edged sword. On one hand, it stands as a beacon of advancement in technology, offering tools that can aid in various fields including, but not limited to, content creation, education, and scientific research. On the other hand, it presents a formidable challenge in the fight against misinformation.

3.1 Generative AI

Nvidia, one of the most known companies producing high-end chips, has reached a market capitalization of $1 trillion, growing by 160% only this year (Wearden, 2023). This phenomenon has shed light on the big rise that artificial intelligence (AI) has undergone in these recent months. Indeed, the famous chipmaker's valuation has increased in response to the great demand of the AI market, in particular after the public release of ChatGPT in November. Apple stocks soared this June after the announcement of Vision Pro headset release and it is only one of the many other companies' surge (e.g., Advanced Micro Devices, AI21Labs, cloud computing services). Today everyone talks about AI, specifically, the debate has now focused on generative AI.

The main goal of artificial intelligence is to create intelligent agents that can act on their own. Generative AI is a subset of AI, the adjective generative stems from the power of generating new content based on inputs. It has been defined as a technology that "*leverages deep learning models to generate human-like content*" (Lim et al., 2023, pag.2).

Deep learning is part of machine learning, the field of computer science that focuses on the creation of models and algorithms capable of learning data patterns without instructions. Deep learning is composed of neural network models, algorithms inspired by human brain neuron's behavior that can learn patterns from massive amounts of data (Zhihan, 2023).

Machine learning encompasses other internal subdivisions, like supervised and unsupervised modeling and generative versus discriminative models. The first classification refers particularly to the data input provided. In the case of labeled data, the model is defined as supervised machine learning since it learns from the past and provides forecasts for the future. Instead, an unsupervised model has no labels and thus it has to extrapolate data features and patterns.

As for the second subdivision, generative models forecast distributions and provide fresh data whereas discriminative models establish probabilities for categorization and decision-making (Zhihan, 2023). In particular, an essential component of generative AI is Natural language processing (NLP), which concentrates on human language to comprehend and produce a variety of material based on linguistic data (Zhihan, 2023). This AI area also encompasses image processing and computer vision.

As explained in the previous definition, generative AI exploits the usage of deep learning and can be approached using two strategies. The first strategy uses conditional probabilities which are at the foundation of the so-called autoregressive models. Transformers and recurrent neural networks are two popular examples of autoregressive models. The second method uses Generative Adversarial Networks (GANs), which are based on adversarial learning. GANs are capable of producing data that is realistic, such as sounds and images. The peculiarity of this model is that the two players, the generator and discriminator, compete during training. While the discriminator tries to tell the difference between actual and created data, the generator aims to provide data that is realistic, lowering the difference between the input and output of the generator. Furthermore, a new category of models entails deep learning neural networks, models capable of extracting features from data and providing forecasts.

The generation of new content in artificial intelligence can vary across different spheres: image generation, code generation, audio generation, and large language models (LLM). The latter has been part of the public debate after the release of OpenAI's ChatGPT which triggered the reaction of many important IT players in the market. Ever since its release in November 2022, high-level competitors like Google Bard and Anthropic's Claude entered the LLM market. LLMs are based on transformer architecture and generate text choosing the next word according to which one has the highest probability of occurring. Of course, the goal of a LLM is to resemble human behavior as much as possible. The progress that has been made is so exceptional that the public debate has been focusing again on the known singularity which generally identifies a specific point in time when AI will overcome humans. How far are we from AI singularity?

3.1.1 GAN

Generative Adversarial Networks (GAN) are machine learning models structured upon an adversarial process, meaning that GAN is a combination of two models: the generator (G) and the discriminator (D). The generator model has the task of generating data as similar as possible to the input ones starting from a noise vector. Thus, its objective is to minimize the distance between real data and the "fake data" it creates. On the contrary, the discriminator needs to distinguish between data created by the generator and real ones: it has to estimate the probability that a sample belongs to real data. Generally, G needs to replicate input data distribution generating new data while D computes the probability that a random sample has been taken from the training data rather than G (Zhou et al., 2018).

The two models are in competition, which means that the generator tries to deceive the discriminator in classifying data and the second tries to avoid getting cheated. This is an example of a zero-sum game where the only way to win is to make the other party lose.

Specifically, the structure of a generative adversarial network can be seen in Figure 3 (Trevisan de Souza et al., 2023). The training starts from the generator which takes into input a noise vector z taken from p_z distribution, representing the latent space Z (Trevisan de Souza et al., 2023). Based on vector z, the generator

G creates synthetic data [1] that is then inserted into the discriminator. However, also original data are integrated into the discriminator. At this point, D has two inputs: the fake data generated by G and the real ones.

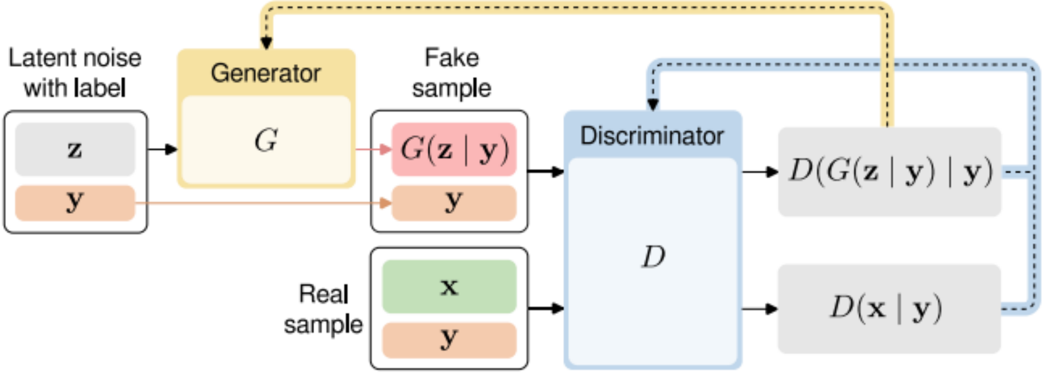

Latent noise with label
z
y
Generator
G
Fake sample
$G(\mathbf{z} \mid \mathbf{y})$
y
Real sample
x
y
Discriminator
D
$D(G(\mathbf{z} \mid \mathbf{y}) \mid \mathbf{y})$
$D(\mathbf{x} \mid \mathbf{y})$

Figure 3- GAN structure (Trevisan de Souza et al., 2023)

The discriminator's task to binarily classify the input into 0 and 1 (e.g. 0 for fake news and 1 for real news) also starts the backpropagation, an algorithm that adjusts model's parameter in order to minimize the prediction error, both for the discriminator error and for the generator one (for the purpose of simplicity L^G will refer to the loss of the generator while L^D to the one of the discriminator). Indeed, loss functions for the two players in the model are structured in different ways.

L^G has to take into account at least these two different components (Zhou et al., 2018):

1. the maximization of the error of the discriminator, therefore minimizing the distance between the fake data and 1. This implies that also fake data would be considered true.
2. Minimization of the difference between generator forecast and real data so that it is much more difficult for D to distinguish truth from false.

As for the discriminator, the loss function contains:

- Distance between real data and 1. The lower this distance, the more performing the model.
- Distance between fake and 0.

The two players are in a constant fight between each other; an ideal GAN is a model in which the discriminator cannot tell the difference between real and false data.

3.1.2 Autoregressive models: Neural Networks

As the name suggests, the structure of a neural network (NN) takes inspiration from the workings of the human brain, in particular brain neurons. The first trainable neural network was built by the psychologist Frank Rosenblatt in 1958 and was called perceptron. According to Rosenblatt, to understand the functioning of human thinking or simply recall, one has to study the "code or wiring diagram" (Rosenblatt, 1958, pag.386) of the nervous system and reconstruct the sensory patterns, how responses to stimuli are created (Rosenblatt,

[1] Artificially generated data created based on real data

1958). Rosenblatt (1958) calls researchers of this field “coded memory theorists” and highlights the similarity between a neuron and simple computer on-off units.

The basic idea behind this model is that the brain is highly complex, and non-linear information is processed in a parallel way, consequently, a neural network should reflect the same workings (Prieto et al., 2016). Therefore, the main features of a NN reported by Prieto et al. (2016) are:

- Big, interconnected networks: a network is composed of units called neurons that are connected among each other and exchange information
- Parallel processing: computations are executed in a distributed manner where units operate independently
- Non-linearity: this feature implies complex relationships that cannot be analyzed linearly.
- Global interconnection
- Self-organization
- Rapid computing
- Learning process helps the model to adapt to the data

The organization of a perceptron is composed of three different types of units: S-points, A-units, and R-units. The first type of unit is the one that receives the input and perceives it with some predefined amplitude that can be an all-or-nothing strategy for some models or different degree according to the stimulus intensity (Rosenblatt, 1958). These signals are then passed to the A-units step distinguished into two parts: Projection area and Association area. In the first area, each cell is connected to multiple S-points, recreating a “projection” of the input. Then we have the association area the cells in this area just receive randomly some fibers from the projections area cells. These units then stimulate a response out of the source set (starting A-units) and then receive feedback. The response step can be divided into two parts: the predominant phase is the step in which some A-units generate a response, but no R-units are active, then the second phase triggers the response of a single R-unit, the dominant response, blocking any other possible activity. Rosenblatt’s model assumes that in order to activate A-units, the input should reach a certain θ threshold.

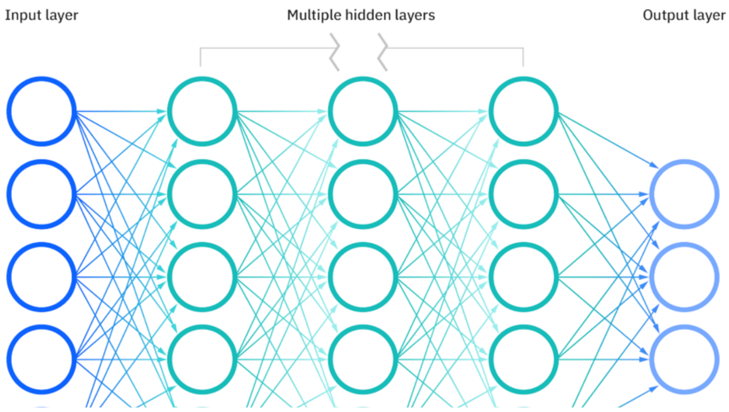


Figure 4- Structure of an Artificial Neural Network (IBM, “What are neural networks?”)

The A-units step lays the foundation of the neuron, for this reason, the perceptron is considered as a neural network with n number of inputs and only one neuron and one output (generated by the dominant response). A neural network is a collection of interconnected neurons that enable the brain to perform cognitive tasks. Similarly, an artificial neural network resembles the functioning of a real brain and consists of nodes connected through mathematical operations where each node represents a place in which computation occurs. The network is composed of three types of layers: input, hidden, and output layers. A more complex structure can contain multiple hidden layers and it is not necessary for all three layers to have a single node. The nodes in these layers are connected to each other through mathematical operations, resembling axons and dendrites. By combining the input to a set of coefficients, called weights, it is possible to assign some degree of relevance to the specific input. Then this information passes through an activation function which turns data into the "activation value" (identifies how much of the neuron gets activated in response to some input) (Garg, 2020). As in the Rosenblatt model, after the activation function, the output is compared to a threshold: only in case it overcomes the threshold then the node activates. Some of the most used functions are the sigmoid, Rectified Linear Unit (ReLU) or Hyperbolic tangent (tanh). The same activation rule spreads across the layers returning an output, this process takes the name of forward propagation. This structure forms the basic framework of a neural network (Garg, 2020). At the beginning of the training, weights and threshold are randomly chosen, then as the activation function spreads the output achieved is improved by adjusting the previous parameters. The result is then compared to the real output and a loss function is computed (difference between real and predicted output). The aim of the model is to minimize the loss and, in order to do so, the model adjusts weights and biases trying to get to a minimum. This process that goes from output to input is called backpropagation (IBM,).

Instead, when referring to deep neural networks, the structure of the model presents multiple hidden layers, conventionally more than 3. Deep NNs are capable of dealing with highly complex information since each layer of nodes is trained on a different combination of features (feature hierarchy). Neural networks can be classified into three different types:

1. Perceptron described by Rosenblatt (1958);
2. Artificial Neural Networks (ANNs) can be also called multi-layer perceptron and are the conventional NNs described above;
3. Convolutional Neural Network (CNN) similar to ANN, they are particularly useful for extracting patterns and features from input data, and they are mostly employed in image recognition;
4. Recurrent Neural Network (RNN) differently from the previous ones, these models can retain a memory of the previous observations for each input and use it to elaborate a prediction.

The last type of the NN is used for the construction of a large language model (LLM).

3.2 Large Language Models

A large language model (LLM) is a type of machine learning model that can carry out a variety of natural language processing (NLP) activities, including translating languages and generating texts as well as classifying texts, and answering questions in a natural way. But what is NLP?

NLP is a branch of artificial intelligence that aims to provide computers with human-like comprehension and interpretation of spoken and written language (IBM). Indeed, this field is the ensemble of various branches like linguistics, machine learning, deep learning, and statistics. NLP's objective is supremely challenging since human language is filled with ambiguities and, using mathematics and statistics as means to give word meaning and context to a computer, requires having a detailed background on several topics, like the ones previously cited. But why natural language? The adjective “natural” has been added to differ language models like Python or Java from natural languages like English, and Spanish (Russell, S. & Norvig, P., 1962-. (2010)). Some of the main NLP tasks are (IBM, “What is natural language processing (NLP)?”):

- Speech recognition: consists in converting voice into text data.
- Sentiment analysis: extraction of subjective qualities like emotions from text.
- Part-of-speech tagging: determining the grammatical part of a word in a speech.
- Natural language generation: it consists in putting structured information into human language.

Computational linguistics, statistical modeling, machine learning, and deep learning are just a few of the technologies that NLP combines. These methods provide computers the ability to understand the intended meaning and feeling of human language, whether it be spoken or written, and to process it.

Precisely, large language models (LLM) are deep neural networks trained on millions of parameters and are built on a Transformers structure. The name of LLM provides us with hints about the characteristics of these models. The adjective large identifies the size of the dataset and parameters which is massive. For this reason, these models can be applied to a large number of tasks and only a limited number of companies have the computing resources to build them.

In order to make these models available to everyone, there have been introduced pre-trained and fine-tuned models that are already trained and can be applied for general purposes.

Instead, a language model can be defined as a model that computes the probability of word sequences, which means that it predicts the next word in a sequence (Zhao et al., 2023). A language model is usually composed of four things (Thiriet, 2023):

1. Number of parameters, patterns of input data
2. Size of dataset, input data can be subdivided into tokens (either single words or pieces of text)
3. Computing performance
4. Network architecture (all LLMs are based on the Transformers architecture)

Language modeling has 4 main areas of development: statistical, neural, pretrained, and large language models (Zhao et al., 2023).

The statistical language models are based on statistics, in particular on Markov chains. Technically, they consist in predicting the next word based on the context, a fixed n-word vector called n-gram. The second type of language model is the neural LM which uses neural networks to shape the probability of word sequences, a good example of this type of LM is the RNN. The key idea behind this model is that predictions are conditioned by multiple features of the context, represented by distributed word vectors. The neural network approach is indeed the one used in Natural Language Processing (NLP) tasks that will be discussed in depth in the following chapters.

Pre-trained models incorporate models trained on large-scale data by someone else to solve a similar problem and they are employed usually in NLP problems. Finally, the focus of this discussion is the last type of language model: the large language model. As anticipated, LLMs are trained on millions of parameters, and this is the reason why they perform very well in solving complex tasks. Some of the most well-known examples of large language models include GPT-3, BERT, and T5. These models can be used for a variety of tasks, such as language translation, sentiment analysis, and chatbots. During training, the model learns to recognize patterns and relationships between words and phrases in the text data, which enables it to generate coherent and contextually relevant language. However, the sheer size and complexity of these models also pose several challenges, such as high computational requirements and potential ethical concerns surrounding the use of these models for potentially malicious purposes. Despite these challenges, large language models have already made a significant impact on various industries and are poised to play an even bigger role in shaping the future of human-machine interaction.

As Zhao et al. (2023) state, these models have existed for years but have really gained a lot of media attention only these days, right after the release of ChatGPT. One of the key questions posed by the researcher is why such LLM capabilities have been emerging only now: what are the real advantages of these language models over pre-trained models which have shown incredible results too? Another point of discussion is their difficulty, training such a model requires significant effort due to the massive computational costs and resources. Furthermore, the biggest obstacle of language models is the similarity to human language regarding subjective traits. For instance, when expressing personal opinions on some facts, AI can easily be recognized. According to Zhao et al. (2023), large language models have the subsequent abilities:

1. In-context learning partially represents the power of LLMs in text generation in response to queries. GPT models are considered the first-movers for this feature.
2. Instruction-following refers to the capability of task execution, LLMs such as LaMDA-PT, released by Google in 2021.
3. Step-by-step reasoning, by employing chain-of-thought (CoT) prompting, larger language models can tackle any task by utilizing intermediate reasoning steps to arrive at the final answer.

Furthermore, Zhao et al. (2023) identify the reasons why these models got so famous all at once. One of the first reasons is the scaling: as the data size increases, the capacity of the model widens; this phenomenon also fosters an efficient computer resource allocation. The latter is significant for the training of a successful model. Indeed, LLMs usually require distributed training algorithms are necessary, which involve the use of parallel strategies. In addition, as anticipated, LLMs possess potential abilities as general-purpose task solvers. However, these abilities may not be explicitly demonstrated when the models are engaged in specific tasks. To address this, suitable task instructions or in-context learning strategies can be designed to elicit and bring forth these abilities, just like the one previously mentioned, i.e., chain-of-thought prompting, and instruction tuning. As for the limitations, Zhao et al. (2023) reveal that their knowledge is constrained by the information present in the pre-training data, making it challenging to capture up-to-date information due to the fact that LLMs are trained on a sequence of time-constrained information. Certainly, this results in missing information. The last point highlighted by the research is data quality. How accurate are the data provided? LLMs are trained on diverse datasets that contain both high-quality and low-quality data, which can result in the generation of toxic, biased, or harmful content. Therefore, it becomes crucial to align LLMs with human values, ensuring that their outputs are helpful, honest, and harmless.

Two important aspects that contributed to the success of GPT models are Transformer language models with the encoder-decoder structure and scalability. These developments have been extremely important to the development of GPT models.

3.2.1 Transformers

Language modeling is based on the study of the probability of word sequences. In machine learning, sequence modeling has always been addressed through recurrent neural networks, which are capable of memorizing previous observations and patterns. However, RNNs have strong limitations in parallelization and long sequences, therefore computer scientists have focused on finding alternative ways to expand sequence modeling's state of art. Several steps forward were made by combining the attention mechanisms with the RNN structure.

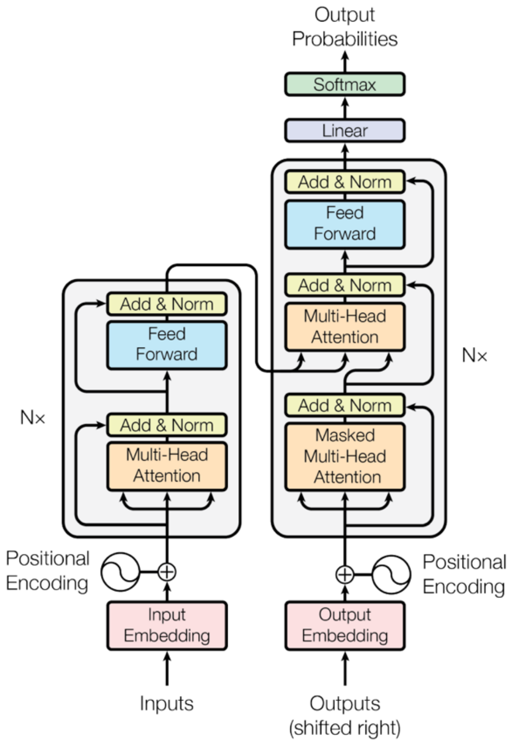


Figure 5- Transformers Structure (Vaswani et al., 2017)

The attention mechanism was first explained by Bahdanau et al. (2014) for neural machine translation which usually relies on the presence of two components: an encoder for the source sentence and a decoder for the target sentence. Using the input sequence as a starting point, the model's encoder creates a contextual representation of the sequence. The decoder then receives this context as input and produces the output sequence. This framework's adaptability enables the choice of several neural networks as the encoder and decoder depending on the precise job at hand. The main job of the encoder is to take the input image and extract the required information.

However, the requirement to compress all pertinent data from a source sentence into a fixed-length vector is one drawback of the encoder-decoder strategy since it makes the training computationally expensive. As a solution to this problem, Bahdanau et al. (2014) proposed an extension of the encoder-decoder model, later described as the attention mechanism.

Taking as an example the research by Bahdanau et al. (2014), their encoder-decoder architecture is built using a RNN in both encoder and decoder. The encoder takes as input a sequence of vectors composed of tokens (pieces of words in a sentence) and all of the vectors are then combined into one vector c, also known as context vector, which is a weighted average of the annotations, elements containing information about the input sequence around the i-th word (Bahdanau et al., 2014). What the encoder does is take c and all the predicted words by the encoder $\{ y_1, \ldots, y_{t'-1}\}$ and compute the product of each word, given the previous ones times the context vector:

$$p(y) = \prod_{t=1}^{T} p(y_t \,|\{y_1, \dots, y_{t'-1}\}, c)$$

In this way, it is not necessary for the encoder to include all of the information in a fixed-length vector since the decoder is embedded with an attention mechanism (Bahdanau et al., 2014).

The transformer model relies entirely on the attention mechanism theory, in particular self-attention and encoder-decoder structure (Figure 5).

In the transformer model, both encoder (left in Figure 5) and decoder (right sequence in Figure 5) are composed of two types of layers, multi-head attention and feed-forward network. Encoders and decoders are stacked on top of each other in equal numbers (the number is chosen by the user as a hyperparameter of the model). Multi-head attention layers are based on the attention mechanism, but the real innovation in these types of models is the introduction of self-attention.

Self-attention is a process that allows one to capture the most relevant words connected to a specific source word, providing context to a specific input. The difference with attention is that the latter actually focuses on the other sequences when making predictions. Instead, self-attention looks at the sequence that is currently encoding (Wydmanski, 2022). This type of attention is much more diffused since its application benefits a lot of natural language tasks like question answering, and text generation.

The multi-head attention layer takes as input a vector x which is then decomposed into three sub-vectors: queries, keys, and values (Kortschak, 2020). In order to build these three elements, the vector x is multiplied by the three weight matrices W_q, W_k , and W_v, respectively representing queries, keys and values of the sentence. The query stands for the word to calculate the attention for; the key instead is useful for mapping the query to the value, it works as a sort of indexing of the sequence's words. Finally, the value represents the word in a sequence.

The function of attention is composed of all three components. The function works in this way: each key corresponds to a value; when introducing a query, the latter is multiplied in a dot product with the K^T. Applying the softmax[2] function is equivalent to finding out which query has the highest score and, by multiplying by V, the process ends up selecting a value. In this way, it is possible to determine how much each word weighs in the self-attention vector. The first attention layer in the encoder in Figure 5 discovers data features and builds the mapping between keys and values. Instead, the first attention layer in the decoder builds the queries. The higher the score, the more connected that word is to the query. The entire formula for the attention function is reported below (Vaswani et al., 2017):

$$Attention(Q, K, V) = softmax\left(\frac{Q * K^T}{\sqrt{d_k}}\right) * V$$

where Q, K, and V respectively represent Queries, Keys, and Values.

[2] Unbounded function that converts values into values between 0 and 1, so that they all sum up to 1. Sometimes they can be intended as probabilities.

As the process iterates, the weight matrices change and at each repetition, a new vector, technically called head, is generated (that is why the layer is called multi-head). As Kortschak (2020) underlines, each head brings about information about different features of a sequence and at the end of the layer, they are all combined into one. An important consideration of the transformer model is that computation is parallelized and thus, makes the whole process fast.

In addition, as one can notice, the self-attention layer is partially different in the decoder and one of the layers is called masked multi-head self-attention, this occurs because some elements in the decoder input are hidden, "masked". Since language decoding a sentence consists in predicting correctly which word comes next in a sentence, by masking some elements the decoder has the challenge of guessing the correct sequence and consequently building the query.

As for the feed-forward layer, both in the encoder and in the decoder, the structure of this layer simply applies two linear transformations with a ReLU[3] activation function.

The real advantage of this model resides in parallel computing and no sequentiality is given as input, it is essential to have a positional encoding in order to indicate the position of the elements. In the paper by Vaswani et al. (2017), the positions are communicated by using sine and cosine functions.

Furthermore, after each sublayer, there is a normalization layer that allows to normalize the results of each layer: indeed, in each layer output of the encoder and decoder are summed up and adding this layer helps in constraining the results in a specific interval.

The last step of the model shown in Figure 5 is a normalization layer followed by a softmax function. The output returned are probabilities for each word, the highest probability belongs to the word predicted by the process.

3.2.2 Scaling Laws

As Zhao et al. (2023) states, one of the most important properties of language modeling is the scaling. Scaling laws in language modeling are influenced by the power law concept which basically describes how one variable changes as a power of another variable. Basically, every time two variables are log-log plotted against each other and their relationship is linear, their distribution is a power law, corresponding to the subsequent function:

$$Y = k\,X^{\alpha}$$

where Y and X are the two variables.

What happens in language modeling is that model's loss diminishes following the power law compared to various factors. Moreover, the higher the number of the variables, the higher the computational power required. Here comes into play parallel training which splits the model across the GPUs to speed up the process. Of course, distributed computing implies batch training: choosing the right batch size is a tradeoff since large batch size worsens performance while small one increases computational power (Wolfe, 2022).

[3] It stands for Rectified Linear Unit and returns the input when the value is positive while 0 for any other value.

Given these premises, in order to study the scaling property of a language model, Kaplan et al. (2020) researched how these factors below influence change in performance:

- Model size
- Dataset size
- Shape
- Context length
- Batch size

Study results reveal that the performance of a transformer model is highly influenced by the scale and not the model shape. In Kaplan et al. (2020) model scale is composed by the number of parameters N (without considering embeddings[4]), the size of the training D and C, the computation power used in the training. Instead, the architectural hyperparameters seem to have a mild effect on the model performance. However, when embedding parameters are included, the number of layers as well as the parameter count have a significant impact on the model's performance. Increasing parameter count (N) and model dimension (D) strongly improves performance. However, if we increase one of these elements while maintaining the other constant, we eventually reach a point where performance improvement is no longer as noticeable (Kaplan et al., 2020). Indeed, by looking at figure 6 it is possible to notice that for high levels of N in combination with D, there is a power law. Instead, for small values of D, as N gets bigger, so does the model's loss.

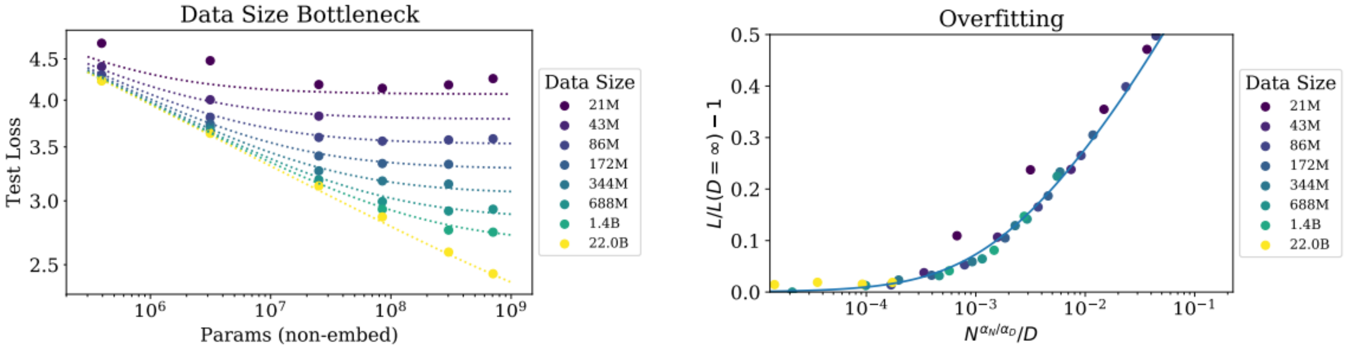


Figure 6- How model's loss changes in response to increases in N and D (Kaplan et al., 2020)

In addition, the computing power available helps in assessing how many batches to create and their size: the higher, the more performing.

Finally, contrary to what can be expected, good language models are not influenced by the variables in the transformer architecture but by the model scale, so its dataset, computing power, and parameters. The resulting loss function proposed by Kaplan et al. (2020):

$$L(N,D) = \left[\left(\frac{N_c}{N}\right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D}\right]^{\alpha_D}$$

[4] An embedding maps a discrete variable into a vector of continuous number, i.e. weights in a neural network

Where α corresponds to the power law exponent, C to the computing budget, N to the parameters count and D to the model dimension. Scaling results by Kaplan et al. (2020) confirm why the new gpt 3.5 model released in November 2022 has quickly conquered the market.

3.3 The LLM society uses every day: *ChatGPT*

In recent years, AI technology, in particular AI chatbots, has drawn a lot of interest. One of the largest language models accessible, ChatGPT stands out for its remarkable technology, based on gpt 3.5 and featuring 175 billion parameters.

OpenAI, a famous research center specializing on AI development, was established in 2015. Their improved language model, ChatGPT, provides conversational answers, error recognition, error learning, and assumption challenging. Initially challenged for its factual correctness, it rose to prominence thanks to its capacity to offer thorough explanations. It simulates interpersonal communication and has a range of verbal responses. Due to ChatGPT's AI capabilities, it is now used in a variety of industries in addition to online customer service, including healthcare, software development, content production, language translation, and customer service.

On November 30, 2022, ChatGPT's inaugural release, which contained a demo version, marked the birth date. On social media, the chatbot soon attracted a lot of interest as users shared stunning examples of its talents. These instances included writing tales, helping with vacation arrangements, and even programming computer programs. Surprisingly, the system attracted over one million users in just five days of launch, demonstrating its appeal and influence.

OpenAI was founded by Sam Altman, Greg Brockman, Elon Musk, Ilya Sutskever, Wojciech Zaremba, and John Schulman in December 2015 (Marr, 2023). Their combined knowledge of software engineering, machine learning, and technology entrepreneurship served as the cornerstone for a company whose mission is to advance artificial intelligence for the benefit of humanity. Elon Musk is no longer participating in OpenAI, but Sam Altman currently is the company's CEO. OpenAI is now valued at $29 billion due to its remarkable success. The business has raised $11.3 billion in total through seven investment rounds. Microsoft and OpenAI increased their collaboration in January, with Microsoft contributing a sizeable multibillion-dollar investment to promote AI on a worldwide level.

Based on a transformer architecture, the gpt series was launched in June 2018 with the GPT-1, text-generation model trained with 117 million parameters (Marr, 2023). The new gpt, GPT-2 was introduced in February 2019 with a substantial increase in the parameters used- around 1.5 billion. It generated cohesive, multi-paragraph content, displaying amazing text generating abilities. It was first withheld from the general public due to worries about potential abuse then slowly made available later in November 2019. With the introduction of GPT-3, people were finally able to engage with ChatGPT and understand the potential benefits of the technology. People may ask questions and get thorough, helpful answers, demonstrating how large language models (LLMs) can be revolutionary.

The most recent version, GPT-4, maintains the upward trend. It adds improvements such greater model alignment, a lower chance of producing objectionable material, enhanced factual correctness, better steerability to adjust to user demands, and real-time internet access for searches (Marr, 2023).

There have been several important developments and milestones in ChatGPT's progress, including (Ray, 2023):

i. Transformers architecture which facilitated the creation of effective and scalable models
ii. The launch of the GPT series, which demonstrated how AI language models may be used for a variety of tasks, including text creation, translation, and summarization.
iii. The introduction of ChatGPT represents the triumph of the gpt-series evolution, with incredible accuracy and versality improvements.

However, as much as technology demonstrated several advances, there are still a number of obstacles and restrictions that must be overcome. For instance, according to Ray (2023), conversational AI models sometimes have trouble preserving context. More logical and pertinent replies would arise from improving ChatGPT's capacity to track and handle context. Of course, this usually degenerates into ambiguity, especially if the query is easily misunderstandable and logical reasoning is often lacking. Subjective responses to queries are also problematic, developing ChatGPT's ability to recognize and respond to users' emotions can enhance its communication effectiveness and create a more empathetic user experience, thus personalizing it (Ray, 2023).

Furthermore, ChatGPT can be subject to malicious people and for this reason it has to be improved to reduce the possibility of producing offensive, prejudiced, or improper content (Ray, 2023). It is vital to continue working on the model architecture, training data, and monitoring procedures.

As anticipated, the peculiar structure of ChatGPT 3.5 released in November relies on a Transformer's architecture. The encoder-decoder functions combined with the self-attention layer allow the model to correctly detect the context of a sentence and the relevance of a certain token compared to the others.

ChatGPT's intelligence and performance on numerous tests have recently been evaluated by research. One important research concluded that GPT-3, the forerunner of ChatGPT, would fall into the 99.9th percentile with an IQ of 150 (Ray, 2023). Furthermore, it had comparable outcomes on the Raven's ability test, demonstrating its cognitive capacities. Ray (2023) emphasizes that GPT-3.5 has proven its competency by succeeding in tests including the US bar exam, CPA, and US medical licensure exam.

The seamless integration of AI into our daily lives improves productivity, creativity, and communication with each milestone. These developments open the door to a fascinating future in which AI will play a key role.

3.3 LLMs and fake news

Are LLMs capable of making mistakes? These models are trained on massive amounts of text data and are capable of generating human-like language responses. As journalists from The New York Times underline, these kinds of models cannot distinguish reality from falsehood and for those who read the "danger is that you can't tell when it's wrong unless you already know the answer" (Hsu, T. & Thompson, S.A., 2023).

Gravel et al. (2023) asked some medical questions to ChatGPT and looked for the references in the answer. As a surprise, most of the sources were fake, created by the deep learning model.

Large Language Models (LLMs) have completely changed how misinformation operations are conducted. In the past, teams of specialists had to put in a lot of work to create political statements that would appeal to particular populations. But LLMs have made this procedure all but obsolete. They have the capacity to produce content that fits into a certain story or appeals to particular emotions, automating the mass production of targeted and compelling information (Mensier, 2023). According to Mensier (2023), the threat posed by misinformation campaigns is increased when Large Language Models (LLMs) are combined with other fundamental models like Stable Diffusion or Midjourney. These extra models can produce incredibly lifelike photos and films, which makes them an effective weapon for disseminating misleading information. AI may dramatically increase the effect of misinformation operations by creating supposedly legitimate multimedia content, including articles accompanied by fake images, making them harder to refute.

The Deepfake[5] video showing President Volodymyr Zelensky caving in on social media is an illustration of the disruptive potential of these models. Due to its importance, this specific instance was simple to refute, but it still shows the disruptive potential of big transformer-based models used in combination. Another notable incident occurred when US Senator Richard Blumenthal presented an AI-generated audio tape of himself giving his opening comments during a Senate hearing on artificial intelligence. This rally emphasized the possible repercussions of LLMs, such as GPT-4, accepting Vladimir Putin's rule or offering approbation of Ukraine's capitulation (Mensier, 2023).

Fighting disinformation is essential for good government and public health. However, given the abundance of disinformation on social media sites, manually detecting false material is a time-consuming operation that cannot be scaled. In comparison to manual efforts, artificial intelligence (AI) approaches have been suggested as a potential and scalable option for identifying false information (Zhou et al., 2023). It's crucial to remember that AI methods themselves might be used to spread false information. One example of AI technology that might contribute to the development of disinformation is Large Language Models (LLMs). These machine learning algorithms can detect, predict, and synthesize human language based on huge databases of human-written information (Zhou et al., 2023).

As Zhou et al.(2023) states this diffused scalable AI technology opens up new opportunities for the spreading of fake news on a massive scale. Meta’s new LLM Galactica was closed after three days from the release. The model was destined to scientists to assist them in academic paper’s summary, scientific code and so on but was found unable to distinguish fact from fiction. Users found that Galactica would create fictitious papers, sometimes even attributing them to actual authors, and create wiki entries on odd themes like the history of bears in space alongside subjects like protein complexes and the speed of light (Douglas Heaven, 2022). When it features space bears, it's simple to recognize fiction, but when it has to do with topics that consumers might not be familiar with, it gets trickier. OpenAI researchers pointed out the trasnformer limitation in 2020 after

[5] Deep learning technique that creates fake convincing images, audio or videos.

the release of GPT-3 by highlighting how the model can actually contradict itself and lack of coherence on long texts (Brown et al., 2020).

For this reason they propose a framework for identifying possible risks connected to the use of Large Language Models. According to OpenAI researchers, threat actors (those who risk misusing LLMs) can be differentiated based on skills and resources at disposal. "Advanced persistent threats" (APTs) are the agents with the highest skills and resources (Brown et al., 2020, pag.10). In particular the paper suggests how large language models actually incentivize and, to some extent, promote misuse of such technology. Indeed, two of the most important incentives are scalability and ease of deployment, since tnfrastructure stability has a big impact on adoption.

Furthermore, the research pursued by the scientists emphasize the dangers of the models: the task of the paper was to classify human-generated text from AI-generated one. What was found was the increasing inability of people to distinguish the two as the model size grew and, consequently, accuracy augmented. Human detection has been described as "close to chance" in this case.

3.3.1 AI-generated text and misinformation

Zhou et al. (2023) analyzes the semantics trying to establish a pattern in the data generated by large language models. The paper employs the linguistic inquiry and word count method to analyze misinformation generated by AI and considers various attributes (Zhou et al., 2023):

1. *Language styles*: taking into account clout or self-centered expressions frequently signals a person's relative social standing, self-assurance, and leadership. Authentic speech demonstrates a freer, more uncontrolled use of words. The level of emotion expressed in communications is referred to as the emotional tone.
2. *Informal attributes*: the usage of casual and colloquial terms frequently heard in regular speech is referred to as using informal language.
3. *Affective attributes*: positive emotions like "good" and "love," as well as negative emotions like feelings of worry, wrath, and grief, are all included in the variety of emotional expressions that language may transmit.
4. *Cognitive attributes*: mental operations that make up human cognition, such as developing insights, comprehending causality, identifying disparities, drawing provisional conclusions, obtaining certainty, and differentiating.
5. *Perceptive attributes*: the capacity or mechanism to perceive something using the senses.
6. *Drives attributes*: people's attempts to achieve particular goals, whether they do so out of a desire for social acceptance, professional success, authority, rewards, or risk avoidance.

Results reveal significant differences between AI- and human- generated text. The first real difference is in the communication styles from those produced by humans, including the ability to change vocabulary when making news vs postings. Particularly, as it can be expected, AI-generated news has less of an emotional tone and more analytical processing and authenticity keywords than human-created news. However, there is a sort of amplification of the emotions which makes the responses given by AI more emotionally powerful with the aim of capturing more attention.

Pan et al. (2023) experiments in “On the Risk of Misinformation Pollution with Large Language Models” focus on determining if intentional misinformation pollution can manipulate QA systems to generate false answers that align with the malicious actor's intentions. By including misinformation in the corpus, performance of QA dropped by anywhere between 14% and 54%. This emphasizes the susceptibility of current ODQA systems to disinformation pollution, which might result from malevolent actors' harmful assaults or language models' accidental hallucinations. Then, according to the researcher, questions that lack substantial justification are more prone to manipulation. Indeed, there was a more significant decline in system performance on the CovidNews test set since this phenomenon generally lacks the same level of informative depth like any other topic. This specific result is in concordance with the logic behind the transformers architecture. In the case of Covid or any other atypical fact, the probability of generating a false word is much higher and the tendency of people in believing such news is also bigger if there is no previous knowledge about that topic.

It is clear that misinformation potential will keep on increasing with the introduction of AI and the need of finding ways to shield society from chaos is always bigger. As Chakraborty et al. (2023) underlines it is also becoming more challenging distinguishing human generated text from AI generated one. As the link between AI and human becomes stronger and starts to affect every part of individual’s daily lives and, as Ruffo et al. (2023) confirms, the only scalable way to combat false news is the same element that can amplify it: AI.

4. FACT-CHECKING: STATE OF THE ART

This section analyses the state of the art of fake news detection, illustrating the pervasive hoax epidemic model and the role of fact-checking in mitigating misinformation. The chapter presents an in-depth analysis of News-Content Based checking, providing insights into different approaches such as linguistic and semantic-based, style-based, and knowledge-based strategies. The chapter escalates to the contemporary automatic strategies, underlining the potential of machine learning in detection systems and its current limitations.

4.1 Fake news detection

Existing fact-checking organizations encompass various aspects, such as slow detection, manual effort, and delayed outcomes. The survey conducted by Singhal et al. (2019) on 88 participants reveals that 98% of them do not care about the veracity of the content they read. About 17% of respondents are still unsure about what constitutes "fake news." However, even if individuals do not care about the news accuracy, the majority (91%) think social media is less trustworthy than conventional news sources (Singhal et al., 2019). Furthermore, when being warned about two modalities (for instance text and image), approximately 71% of participants expressed confidence in their ability to distinguish between genuine and fake news.

Therefore, web users must improve their discernment abilities to distinguish between reliable and false news. There are three different approaches for classifying news: news content-based, social context-based, and creator-based (Kondamudi et al., 2023).

1. News-content based are the most common systems used for fake news classification since it is easier to extract features from information. False news reports are usually always spectacular and eye-catching to attract more hits and interest. The news stories may not relate to the headlines at all or may even contradict the main news's assertions of reality. When defining a piece of questionable internet material, it is highly recommended to read the entire article rather than just the highlighted news. In addition, article authors typically use a range of facts to persuade readers, including information from experts, research or analytical data, citations, and even links to relevant news articles. Most false news tackles viewers' worries, anxieties, sympathies, exhilaration, apprehension, and other emotions on purpose. Therefore, internet users should be responsible for assessing the delicate emotional grade while adhering to the sentiments and viewpoints in the material. For instance, if the news makes you angry or downhearted. Additionally, when tragic news items are posted, users of social media must be aware that they are dealing with a delicate subject and continuously question whether the material appears a little too appealing or funny to be genuine.
2. Social context-based: involves studying the context of the news, such as the examination of the source site or the time period of the fact.
3. Creator-based: A rising number of academics think that the best way to spot misleading propaganda is to focus on news origin rather than the exact statements themselves. Online users can recognize a variety of social cues that can assist them in identifying suspect sources and possibly deceptive information. Analyzing the linguistic features of the URL can assist in identifying rogue websites,

(Singhal, 2019) regardless of whether they are well-known or obscure online entities. Red flags might be raised by strange site addresses (such ".com.co") or dubious signs in the URL.

The study will focus in particular on the first type of news classification.

The task of CBFND (Content-Based Fake News Detection) involves using numerous features, which are often categorized into linguistic, grammatical, readability, and sentiment features in the analyzed studies. However, some works include the same feature in different groups (Capuano et al., 2023). In the past, identifying false news sites largely depended on text and user information elements. To enhance fact analysis, certain methods made use of knowledge graphs and entity relation data. Researchers started experimenting with several modalities, such as text and graphics, and integrated them for richer data representation in order to address the difficulty of brief and informal social media data (Singhal et al., 2019).

Even though all the single-modal strategies listed above were able to produce encouraging results, the informal and brief character of social media data makes information extraction difficult. The researchers began experimenting with characteristics taken from other modalities (such as text and picture) and merged them together for richer data representation to get over this constraint, giving origin to the multi-modal detection approaches (Singhal et al., 2019).

Indeed, single-modal false news detection looks at just one kind of feature whereas in the case of multi-modal fake news detection, various feature combinations are examined. Recent research works developed in multimodal detection combine two kinds of feature sets: a mix of visual and textual elements, and user profile and news content elements (Athira et al., 2023). An example of this approach is Spotfake, a model created in 2019 by Singhal et al. Spotfake is a multi-modal detection approach capable of handling textual and visual elements. Paper results show its superiority across neural networks and variational autoencoders, achieving large accuracy gains on Event Adversarial Neural Networks for Multi-Modal Fake
News Detection (EANN) by using a language transformer model and pre-trained ImageNet model.

4.2 Hoax epidemic model and fact checking

As anticipated, there are controversial opinions on fact-checking as an effective solution against fake news. However, as tech flows towards the so-called technological singularity, fact-checking still remains essential in warning individuals against low-credibility and low-accuracy news (Myojung & Nuri, 2020).

Myojung & Nuri (2020) confirm that when users are warned about the potential falsity of news they are less reliant on information content. Indeed, the research also shows that users are more willing to share news when social media are at its highest. The impact of social media metrics on social sharing intentions is, however, overruled when information about fact-checking exposes false information, thus highlighting the effectiveness of fact-checking (Myojung & Nuri, 2020).

The research of Clayton et al. (2020) studies how Facebook’s anti-misinformation campaign impacted the perceived accuracy of readers in 2016 through the usage of automated fact-checking Facebook implemented warnings on posted content and tags containing “Disputed” or “Rated false” text. Warnings have a negligible impact on people believing false information and are thus less effective than the targeted tags. Furthermore, general warnings diminish trust in credible news sources and do not enhance the effectiveness of both tags, employing more targeted tags emerges as a safer approach to mitigate the dissemination of false information without unintended adverse outcomes. (Clayton et al., 2020). Instead, tags demonstrate a good influence on perceived accuracy which confirms the underlying thesis: fact-checking benefits can limit false information-spreading contexts.

At the same time, as cited above in Ruffo et al. (2023), the effect of fact-checking on misinformation is controversial and can impact negatively as much as positively the spread. Indeed, due to the hyper-correction, this solution can actually increase the polarization and give origin to echo chambers.

Furthermore, fact-checking accuracy largely depends on the method used. Of course, the most reliable way to verify news is using professionals to countercheck each piece of information but, as it can be easily imagined, such a method cannot be applied on a large scale and it is very time-consuming (Ruffo et al., 2023). The most common mechanism to detect misinformation is automated fact-checking consisting in the use of algorithms to spot risky information and, even if this is a rather scalable method easily deployable on a mass-level, it can be less accurate and explainable. Recently, a new methodology has been devised for this purpose: crowdsourced fact-checking, which benefits from the help of common users in checking online content. Scalability in this case is much higher than in professional one and accuracy is very high (even if this assumption does not always hold). A mixed strategy approach is the one that Ruffo et al. (2023) identify as "human-in-the-loop". This last proposed methodology uses automation to slim down the list of tweets to be checked and then the human factor finalizes the fact-checking, increasing accuracy.

As previously anticipated, fact-checking dynamics are explained by the hoax epidemic model. Each agent can become one of three states (Tambuscio et al., 2015):

- Susceptible: users who have not been exposed to fact-checking
- Believer: the ones who believe and spread falsehood
- Fact-checker: the ones who spot the fake news and spread the correct information

It is important to underline that the susceptible agent is assumed to be neutral. However, by recalling subjective bias, the neutrality assumption may not hold in most of the cases.

The research analyzes two epidemic models: SIR and SIS, respectively Susceptible-Infected-Recovered and Susceptile-Infected-Susceptible.

Therefore, Tambuscio et al. (2015) research studies three actions:

- Spreading: each agent can change its beliefs through neighbors' opinions, the spreading action occurs either because a susceptible agent turns into a believer or a susceptible becomes a fact-checker. In the first case you spread incorrect information, in the second case verified one.
- Verifying: this behavior only occurs when the believer decides to fact-check with p_{verify}
- Forgetting: can be observed in the SIS model where the agent exposed to the fake news believes it and forgets it, ignoring the spreading

The model allows a believer to become a fact-checker but not the other way around. The probability that news gets shared depends on the number and the type of agents in the neighborhood of a single user. Of course concepts such as polarization and echo chamber reinforce this probability.

According to the hoax model, considering a starting status quo of believers, there exists a state in which the number of susceptible users at infinity stabilizes, which means that it is possible to study when the actual diffusion of misinformation stops and the number of believers tends to zero.

The parameters Tambuscio et al. (2018) add to the model are mainly two: credibility α and forgetting probability p_f. When credibility parameter is low, it means that the users do not believe the fake news and are therefore the group defined by the researcher as skeptical while high values of α correspond to the gullible group. At this point the other parameter comes into play: when the forgetting probability is high misinformation spreads in both groups while in the case of a small p_f, the believer size changes according to the segregation level between the two groups. Concretely, in case of conspiracy theories which present a low forgetting probability, the size of believers is at its highest and the elimination of fake news is a highly difficult task. Indeed, if every agent was allowed to change status, the worst outcome would be converging towards a believers-predominant network, neutralizing the effect of fact-checking. However, as Ruffo et al. (2023) state, fact-checking is always worth the try also in the worst possible situation since it could still create a minority of debunkers.

Therefore, as the consequences of automated fact-checking can benefit as well as damage society, it still remains the only scalable way capable of warning mass society against the possible misleading, manipulated, invented content.

4.3 News-Content Based analysis

News content-based fact checking refers to the process of verifying the accuracy and truthfulness of news articles and information by analyzing their content. Indeed, the observable characteristics of real news, including headlines, body text, photos, and videos, can show biased and prejudiced behavior and act as signs of false news. It takes a variety of analysis to classify false news based on news content, including linguistic and semantic evaluations, knowledge-based assessments, and style-based assessments.

Instead of relying solely on external sources or expert opinions, this approach focuses on examining the language, context, claims, and evidence presented within the news content itself to determine its validity.

News content-based fact checking is valuable because it offers a way to quickly assess the credibility of news articles without solely relying on external sources.

4.3.1 Linguistic and semantic-based approach

This analysis digs deep into the related language forms, patterns, and meanings found in news articles to glean insightful information from the writing. There are several linguistic and semantic-based analysis: expert-based and semantic based (Kondamudi et al., 2023). The former aims at highlighting the writing style of a specific news through the examination of its grammatical structures. Tai et al. (2020) employees “bag of words” and “n-grams” approaches to examine the news structure. A “bag of words” summarizes the information content into collection of individual terms whereas n-grams consists in a sequence of n elements. However, Kondamudi et al. (2023) emphasized that these methods have some limitations: for instance the may lose important information by disregarding word semantics, whereas "n-grams" might result in a large feature space and struggle with unfamiliar terms and this is the reason why some other methods like word2vec and LSTM took over.

On the other hand, semantic-based analysis consists in drawing meaning from text by focusing on the sentence at different levels: from word to context level following the syntax rules (Kondamudi et al., 2023).

Some articles were found to have sensational headlines which were inconsistent with the news’ content, the sentiment analysis falls into this broad category.

4.3.2 Style-based approach

The ability to discern between the writing styles of individual identities is essential for identifying fake news on the internet, and style-based analysis plays a key part in this process. Based on its goals, this analysis technique may be divided into two categories: physical style evaluation and non-physical style assessment (Kondamudi et al., 2023).

The physical style of a text involves identifying substantial physical indicators that can distinguish fake news from genuine news. These cues might reveal the writer's writing style, text organization, and individual traits, encompassing aspects like verb and noun frequency, emotional language usage, and informal expressions. Additionally, detecting suspicious elements in social communication such as URLs, hashtags, comments, and capitalized letters proves valuable for both authorship identification and the evaluation of writing style. Non-physical assessments focus on the abstract qualities of knowledge, including things like message complexity and depth. Researchers Tai et al. (2020) found that those who produce fake news frequently take longer to write and make more mistakes. This makes it possible to identify particular keyboard patterns that reflect their writing style. For instance, false news writers regularly use keys like "backspace" and "delete" while creating fraudulent information (as cited by Tai et al., 2020 in Kondamudi et al., 2023). These methods can help determine the intention behind news content, whether it's meant to deceive or not. However, the use of style-based techniques to create fake news in a distinctive manner (based on beliefs and emotions) raises concerns. To address this, a comprehensive investigation is conducted to systematically identify these unique content types using machine learning techniques.

Hence, according to Kondamudi et al. (2023), the efficiency of systems for style-based false news detection depends on: the capability of communicating text or visuals of the message, performance of the classifier across various formats, example of bogus news and the style-based strategy.

As anticipated, in order to devise a successful method for fact checking, understanding and maintaining the style of the original news is essential. The style of a message is perceived through the textual features as well as the image ones. Textual features commonly used in machine learning frameworks to detect fake news are based on language proficiency levels: lexicon, syntax, discourse, and semantics (as cited by Conroy et al., 2015 in Kondamudi et al., 2023). Some of the techniques used to pursue a textual analysis are bag-of-words (BOW) and part of speech (POS) which can be either nouns or verbs or any other element in a sentence.

Latent textual characteristics are used to encode textual news, which may be done at the level of individual words, phrases, or entire texts.

In classic machine learning systems, these representations, which take the form of vectors that summarize a news story, may be utilized as direct inputs for classifiers like support vector machines to identify false news (Kondamudi et al., 2023).

Furthermore, these structures can be also used into neural networks and transformers structures to identify fake news patterns.

4.3.3 Knowledge-based approach

This method involves leveraging a pre-existing repository of collective human understanding to ascertain the truthfulness of new statements. The primary benefit of this approach is that, beyond providing a label, it can also offer explanatory context.

4.3.3.1 Manual fact checking

Manual fact-checking can be categorized broadly into three main types: Expert-Based Manual Fact-Checking, Expert-Based Fact-Checking Websites, and Crowdsourced Manual Fact-Checking (Kondamudi et al., 2023). The first type of fact checking leverages the help of experts to debunk falsehood. It is simple to use and produces reliable results when the size of the news is restrained, as the amount of material to be analyzed grows, it becomes more expensive and less scalable.

Also, websites provide fact-checking functions based on expert opinions. Examples to note include Hoax-Slayer and PolitiFact. The PolitiFact grading system provides information on the veracity of claims made on particular subjects, aiding in establishing the validity of articles and flagging problems that need more research. However, some of the limitations present in current fact-checking websites include extended detection times, delayed outcomes, and a substantial requirement for manual labor. Consequently, also online consumers need to enhance their capacity to differentiate between authentic and deceptive news.

Finally, Crowdsourced Manual Fact-Checking involves every user to contribute to the process by breaking down big tasks into smaller ones. In this method, regular users function as fact-checkers. Numerous fact-checkers might be enlisted by using different crowdsourcing strategies. Crowdsourced fact-checking offers access to a sizable fake news dataset, but because it depends on a varied group of fact-checkers, there may be issues with accuracy and dependability. It is vital to eliminate untrustworthy sources in order to produce more reliable results.

4.3.3.2 Automatic fact checking

Manual fact-checking is difficult because of the growing amount of freshly created material, especially when it comes to social media platforms. Automated fact-checking procedures have been created to address sustainability challenges, largely depending on machine learning techniques, information retrieval, natural language processing, and network/graph theories. Data can be represented by the combination of multiple elements: subject (Y), predicate (Z) and object (I). As the data is compared to the knowledge then it can be considered true. Indeed, knowledge is established once the information has been confirmed by previously known news, the source truth. The fact-checking process here divides into two subprocesses: fact extraction and fact-checking. It is crucial to first get information from the World Wide Web as basic "facts" that later need to be evaluated. There are two types of knowledge extraction: single-source and fully accessible. Single-source knowledge extraction depends on a single reliable source (such as Wikipedia) to collect pertinent knowledge; of course, in this case the knowledge retrieved is limited. As contrasted with single-source discovery, fully accessible extraction combines data from various sources, producing more complete findings at a reduced cost (Kondamudi et al., 2023). As Kondamudi et al. (2023) highlights, there are some obstacles in obtaining correct information to build a source of truth:

- Redundancy: same elements/figures might be represented differently in a text, for instance the name Barack Obama might be also written as Barack Hussein Obama.
- Invalidity: some truths might be outdated; a good example is Great Britain that exited European Union
- Conflicts: conflicting information or claims create opposing viewpoints
- Unreliability: some sources can contain low-quality and credibility information
- Incompleteness

In the fact-checking step the YZI information needs to be aligned with truth, knowledge base. But what happens in the case in which the data is non-existent in the knowledge base?

Kondamudi et al. (2023) lays out three assumptions that address this issue: closed world, open world and local closed world assumption depending on which the YZI triple can be inaccurate, incorrect, or uncertain.

Knowledge-based and style-based news content analysis is fundamental for a successful development of a machine learning methods to spot fake news. Moreover, as anticipated the construction of a solid pre-existing unbiased knowledge is particularly difficult to obtain.

4.4 Machine Learning: how can AI limit fake news diffusion

As Ruffo et al. (2023) analyzes in their research, automated mechanisms to detect fake news allow to counter-respond fast to misinformation and address the scalability problem. Current AI solutions sacrifice accuracy for the latency of “more precise” solutions. Indeed, human factor could highly enhance algorithm performance, but they require much more effort and time which could be detrimental to society, considering the velocity at which falsehood distributes. Given the millions of news articles shared on social networks daily, it becomes easier for online users to distribute stories with attention-grabbing headlines without thoroughly scrutinizing their accuracy. This is where machine learning come into play.

Most recent studies on the automated detection of false news have concentrated on a specific angle, such Natural Language Processing (NLP) or data mining (Kondamudi et al., 2023). Finding out whether a source of information is reliable is the process of fake news identification. As previously anticipated this process takes the name of fact-checking and, in the case of automated mechanisms, it involves machine learning methods. Unsupervised, semi-supervised, and supervised models are the three categories under which machine learning approaches for data mining fall. In recent years has been a rise in the popularity of deep learning algorithms, notably in the domains of voice and object recognition.

Systems that watch social media accounts or models that track the dissemination of fraudulent material to detect bot or spam profiles are some measures taken to stop the spread of deceit. One of the initial attempts in this regard was made in 2010 by Benevenuto's group (Benevenuto et al., 2010), who utilized a non-linear Support Vector Machine (SVM) with a Radial Basis Function kernel to detect spam accounts based on user behavior attributes. The model successfully identified 70% of spam accounts and 96% of non-spam accounts from a dataset exceeding 1,000 entries.

Additionally, research from 2010 (Chu et al., 2010) titled "Who is tweeting on Twitter: human, bot, or cyborg?" offered a novel categorization scheme to distinguish non- and human accounts. The system's peculiarities were its four parts: an entropy component for detecting automated posting times, a machine learning component (Bayesian classification) for detecting text patterns, an account properties component for detecting bot deviation from the typical distribution of humans, and a decision-maker component (Linear Discriminant Analysis) for making the final call. The system's accuracy in recognizing human users was 96%. Ma et al. (2017) created a more complex model that used propagation trees to describe news in order to track how a message is changed over time by users. Based on structural and linguistic characteristics, the model employed a Propagation Tree Kernel to compute the similarity between rumor trees and non-rumor trees. On two separate datasets, the approach had accuracy rates of between 73 and 75%.

Despite the encouraging results, most of this research focuses on bot detection there is a fundamental downside to this approach: in order to identify a spam account, it must post enough bogus information to build a useful profile. The dissemination of already circulated news cannot be stopped, even though the source can be eliminated once it has been found. Additionally, a fresh bot account may be instantaneously generated to keep the process going indefinitely. Contemporary solutions are more focused on classifying articles primarily based on their content.

According to Khan et al. (2021), deep learning models are often better at detecting false news than standard machine learning models. The conventional machine learning model Naive Bayes obtained 93% accuracy. Two deep learning models, Bi-LSTM and C-LSTM, both reached 95% accuracy on the combined corpus, a substantial increase. These findings imply that deep learning algorithms may detect bogus news more accurately than conventional machine learning techniques.

4.4.1 Supervised and unsupervised learning for misinformation

Kondamudi et al. (2023) presents a mapping of all machine learning techniques both for the supervised and unsupervised ML. The practice of using labeled datasets to train algorithms for precise data categorization and result prediction is known as supervised machine learning, and it is a subfield of artificial intelligence.

For the detection of online hoaxes, frauds, and misleading classification of information, a variety of supervised machine learning techniques, including Logistic Regression, Random Forest, K-nearest Neighbor, Decision Tree, and Support Vector Machine (SVM), have been extensively used in the past (Kondamudi et al., 2023).

In Abdullah-All-Tanvir et al. (2019) cited by Ahmed et al., (2021), machine learning classifiers, such as SVM and Naïve Bayes, were employed and demonstrated superior performance in detecting fake news based on their accuracy. The efficacy of a classifier is greatly influenced by its accuracy. A classifier that is more accurate is thought to be better at spotting bogus news. In Aphiwongsophon & Chongstitvatana (2018) cited by Ahmed et al., (2021), their proposed models were SVM, Naïve Bayes and Neural network. Incredible accuracy was registered by the second. Machine learning classifiers were used by researchers (Reis et al., 2019, cited by Ahmed et al., 2021) to identify fake news, and they were trained using a variety of characteristics. To get reliable findings, they underlined how crucial classifier training is. According to Granik, M. & Mesyura, V. (2017), false news can be identified from Facebook postings with an accuracy of 74% thanks to artificial intelligence, notably the Naive Bayes classifier. They also suggested approaches to boost accuracy even further.

Other methods worth mentioning are the decision trees, logistic regression, random forest, and k-nearest neighbor. The research "Fake News or Truth? The breakthrough method described in "Using Satirical Cues to Identify Potentially Deceptive News" examined elements of satirical news such absurdity, comedy, syntax, negative affect, and punctuation. Their use of TF-IDF and SVM algorithms improved outcomes, yielding an accuracy rate of 82%. This demonstrated the potential efficacy of identifying complex grammatical and syntactic patterns as trustworthy markers of deception (Rubin, Conroy, Cornwell, & Chen, 2017).

Unsupervised learning models are extremely useful and practical for dealing with difficulties in the real world. However, there is little study that focuses especially on identifying the unchecked internet dissemination of fake news (Kondamudi et al., 2023). This is due to the fact that finding the correct labeling is very difficult and finding high quality data is very challenging (Kondamudi et al., 2023).

If getting good results with supervised learning is achievable, applying a fake news detector on unlabeled data is very challenging but also more important since news circulating on social media are not labeled. Indeed, some of the techniques used by Kondamudi et al. (2023) to approach this kind of difficulty are semantic similarity between fake and real news, clustering that can help on determining groups of news to attention, outlier analysis and unsupervised news embedding which relates to the natural language processing (NLP) process by which the sequence of words are translated into vectors.

4.4.2 Deep Learning for misinformation

The proliferation of fake news in online social networks and media has posed significant challenges in ensuring the accuracy and reliability of information. As the problem becomes increasingly complex, researchers have explored various approaches to detect and combat fake news. In recent years, there has been a growing interest in utilizing deep learning techniques, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bidirectional LSTM (BiLSTM), for fake news detection. Moreover, deep learning models, particularly those incorporating attention mechanisms, can effectively handle the dynamic nature of fake news by selectively attending to relevant parts of the text. These attention-based models focus on salient features and weigh them differently, allowing the model to assign more importance to critical information for accurate classification. Recurrent neural networks (RNNs) and long short-term memory (LSTMs), two particular deep learning architectures, have been shown to be efficient in identifying sequential patterns and spotting bogus news (Li et al., 2022b; Van Houdt et al., 2020 cited by Kondamudi et al., 2023).

The dense neural network used by Thota et al. (2018) to classify bogus news obtained a remarkable accuracy of 94.21% by feeding Term Frequency-Inverse Document Frequency (TF-IDF) word representations and conducting some feature engineering into a dense neural network.

Ahmad et al. (2020) proposes an ensemble-based false news classification strategy utilizing many textual features. Although this method produced excellent accuracy ratings, conventional machine learning algorithms like this are prone to false positives.

In a study by Kumar et al. (2020), several approaches were compared, including CNN, LSTM, and attention mechanisms, concluding that a combination of these mechanisms achieved the high accuracy of 88.78%. Nevertheless, optimizing LSTM models proved challenging due to their long training times and sensitivity to initial weight values. The application of NLP also significantly increases training time of algorithms.

Fake news detection has been carried out using a variety of deep learning models, including bidirectional recurrent neural networks (RNNs), gated recurrent units (GRUs), and Transformers (Kondamudi et al., 2023). Khan et al. (2021) states that the greater overall accuracy attained by deep learning models in comparison to traditional models indicates that deep learning models typically outperform traditional machine learning models in the identification of false news. When working with big datasets, this discrepancy becomes more obvious, demonstrating that deep learning models have a propensity to overfit when trained on smaller datasets. Although Naive Bayes (with n-gram) is a conventional model, it exhibits good results in the identification of fake news, matching deep learning models' performance with an accuracy of 93%.

4.4.3 Advanced language models

Advanced pre-trained language models like BERT, ELECTRA, and ELMo are now receiving a lot of attention for a variety of natural language applications, including text categorization. However, very few research have looked at how they may be used to identify false news. BERT stands for Bidirectional Encoder Representations from Transformers and is a is a pre-trained model designed to learn contextual word representations from unlabeled texts. An enhanced variant of BERT called RoBERTa (Robustly optimized BERT method) performs better thanks to training on bigger mini-batch sizes, longer sequences, and more data. Instead, ELMo and ELECTRA are also based on language modeling architecture.

Jwa et al. (cited by Khan et al., 2021) showed an improvement in F-score compared to earlier models when using BERT to identify false news by examining the link between the title and body content of news items. Some research has also focus on the application of language models in combination with deep learning.

Despite having more intricate architectures, these models, when compared to deep learning models, show less overfitting on smaller datasets (Khan et al., 2021). This is due to the fact that all layers of these models—aside from the final classification levels—use pre-trained weights. As a result, they don't need a big dataset to optimize their intricate design. This is indeed the big advantage pre-trained models offer.

The research also reveals that the amount of pre-trained parameters closely correlated with the performance of the transformer-based models.

An interesting approach was the one of Shu et al. (2019) with the model dEFEND comprising a word encoder, a sentence encoder, a user comments encoder, a sentence-comment co-attention layer, and a fake news prediction component. Each sentence related to a specific comment was encoded using a co-attention layer. This layer was so important that accuracy dropped when either co-attention for news content or user comments were excluded, indicating the significance of user comments in guiding fake news detection.

This dEFEND model outperformed others, achieving an accuracy of 90.4% and an F1 score of 0.928 on the PolitiFact dataset, and an accuracy of 80.8% with an F1 score of 0.755 on the GossipCop dataset.

The real benefit of transformer models for fake news classification can be seen in Rai et al. (2022) which proposes a combination of BERT and LSTM (BERT+LSTM) on the FakeNewsNet dataset, which includes PolitiFact and GossipCop datasets. The results of the comparison revealed that the proposed model achieved a maximum accuracy of 88.75%, outperforming the other models. Notably, the model demonstrated an accuracy improvement ranging from a minimum of 1.35% to a maximum of 17.55% when compared to the baseline models on the PolitiFact dataset. Similarly, when evaluated on the GossipCop dataset, the proposed model exhibited an accuracy increase ranging from a minimum of 0.3% to a maximum of 10.5% when contrasted with the baseline models. Indeed, the Politifact dataset accuracy reached 88.75% of accuracy, winning over all the previous models like TCNN-URG, LIWC and simple BERT.

In Khan et al. (2021), all of the pre-trained models beat deep learning and classical machine learning models according to the metrics.

According to Kai et al. (2022), the performance of transformer models on fake news detection overcomes all of the other machine learning models. Some research by Aggarwal et al. (2020) cited in Kai et al. (2022) confirms that a simple fine-tuned BERT reached an accuracy of approximately 97% .

4.5 Limitations of current approaches

While deep learning algorithms show promise for fake news detection, they also come with their own set of limitations. Fake news is a complicated problem that involves many different languages and user-unfamiliar modalities including text, audio, and pictures. In order to improve performance, the use of sequence models for processing lengthy textual features necessitates careful evaluation of feature and classifier selections.

Another potential research domain is the development of models for detecting fake news. These models would combine textual and visual features with correlation analysis to identify patterns in social media news posts that are indicative of fake news (Athira et al., 2023). Athira et al. (2023) research focuses also on an important limitations of deep learning algorithms: explainability. Although the effectiveness of this false news detection technique seems promising, it is unclear how well it can be explained, which casts doubt on the conclusions reached. Newer techniques, in contrast to older detection models, use more data, such as social context, multimedia content, and user details to improve detection accuracy. For this reason, it could be useful to have a multi-class fake news detector that classifies fake news into different classes based on some indicators.

Data quality concerns in the fake news detection field has also been addressed by Capuano et al. (2023). Indeed, the utilization of varied datasets in fake news identification presents obstacles in terms of data collecting, verification, and storage procedures. A defined method is required to guarantee consistency and quality in dataset production. Progress in this field of study is hampered by problems with data sources, verification procedures, and the narrow range of topics covered by databases, notably political themes.

Moreover, according to Capuano et al. (2023), studies already conducted mostly use databases in English or a hybrid of English and another language. It is difficult to evaluate the effectiveness of algorithms and features from a multi-language viewpoint because of this narrow emphasis. Additionally, models must be able to handle the wide range of subjects that are covered by different datasets. It is challenging to judge a model's capacity to generalize across different themes and data sources, such as social media postings or conventional web news, because much research only analyze their models on a single dataset. As a result, there is no unified way to verify sources and all the algorithms proposed focus only on just one type of data source.

Another important key problem is that the social landscape keeps on changing. Therefore, an adaptive continuous learning model is required since the strategies used by individuals who propagate false information are always changing. Such a model might adjust to the shifting patterns and traits of false information. To successfully address this problem, it is essential to create a model that can constantly adapt and learn from new patterns of false news propagation (Capuano et al., 2023).

One of the most important challenges is that Natural Language Processing (NLP) applications used to identify false news are vulnerable to numerous assaults that target machine and deep learning models. The effectiveness of NLP models can be greatly impacted by these attacks, which include the falsification of information, manipulation of subject and object, and the establishment of erroneous causal linkages. Exaggerating or changing particular words results in distortion, which leads to erroneous interpretation. In order to sway the reader's view, the tactic of confusion of cause involves inventing causal relationships between unconnected events or giving only a portion of a tale. Attacks of this nature make it difficult to accurately identify false news using NLP approaches.

In conclusion, Puravain et al. (2023) demonstrate that despite efforts to improve linguistic characteristics by using cutting-edge algorithms like XGBoost or BERT, only few classifications are able to exceed the 90% accuracy barrier. Furthermore, it continues to be difficult to identify and classify the proper language traits for a given situation.

A study from Choudhary & Arora (2021) tries to to identify the content characteristics that influence language-guided aspects. The model extracts syntactic, grammatical, emotive, and readability aspects from particular news stories,. Existing literature served as inspiration for these qualities. Syntactic characteristics are preferred since they are used so frequently. Characteristics including character count, word count, title word count, stop word frequency, number of capitalized words, and keyword density are included in this list of properties. It's crucial to remember that none of these linguistic characteristics directly imply syntax. The results of this study are encouraging, although their interpretability is not very good. Therefore, correct classification and categorization of these traits might aid an educated selection process, leading to a more fruitful study. To summarize, classifications with more than 90% accuracy are still uncommon, and linguistic elements are frequently picked without a good reason—likely because they are common or easy to acquire in popular libraries. We think a careful classification of linguistic characteristics could direct their purposeful selection. Hasty feature selection decisions may lead to methodological ambiguity and make it more difficult to comprehend results.

5. PROPOSED SOLUTION

5.1 Data collection and Exploration

Building a good dataset for fake news detection in automated mechanisms is one of the most difficult tasks. As Capuano et al. (2023) underlines, processes like data collection and verification do not always satisfy data quality requirements.

For this model data was collected via web scraping from Politifact, independent American fact-checking institution, and ranges from 2007 up until September 10th, 2023. Web scraping is the automated process of extracting information from websites, transforming unstructured data of web pages into structured data that can be analyzed, stored, or utilized for various purposes. It involves using specialized software or tools, often referred to as web scrapers or web crawlers, to navigate through the web's HTML (Hypertext Markup Language) structure and retrieve specific pieces of information.

Web scraping works by sending HTTP requests to the target website's servers, requesting the HTML code that constitutes the web page. Once the HTML content is obtained, the web scraper parses through the code to identify the relevant data points, such as text, images, links, or other elements. This process is often facilitated by libraries or tools designed for parsing and extracting information from HTML, such as Beautiful Soup for Python.

Web scrapers can be programmed to follow specific patterns, such as navigating through multiple pages or following links, to collect data from multiple parts of a website or across different websites. They simulate human interaction with web pages, extracting data at a much faster rate than manual extraction. However, it's important to note that web scraping should be done responsibly and ethically, respecting websites' terms of use and not overloading their servers with excessive requests, which could lead to performance issues or legal concerns. In essence, web scraping offers an efficient way to gather information from the vast expanse of the internet, enabling data-driven insights, research, analysis, and automation of various tasks that rely on web-based content.

Web scraping was enacted on Politifact which is a website serving as a credible and authoritative source for fact-checking claims made by politicians and public figures on social media. Its extensive collection of statements, ranging from those that are entirely accurate to those with varying degrees of falsehood, provides a rich and diverse dataset for training and testing fake news detection models. The website's commitment to unbiased reporting and meticulous verification makes it an ideal resource for gathering a diverse range of news content to bolster the dataset's authenticity.

By utilizing Beautiful Soup, the process of scraping news articles from the PolitiFact website was streamlined. This involved parsing the website's HTML structure to identify and extract relevant news articles and their associated metadata, such as the source, context, and truth rating. This meticulously curated information formed the foundation of the fake news detection dataset. PolitiFact utilizes a comprehensive rating system to assess the accuracy and truthfulness of statements and claims made in news articles, speeches, and other sources. This rating system, often referred to as the "Truth-O-Meter," categorizes statements into six distinct ratings, each denoting a different level of accuracy (Politifact, s.d.). These ratings provide readers with a quick and easily understandable assessment of the veracity of the claims being analyzed.

1. True: This rating is assigned when the statement or claim being fact-checked is completely accurate and backed by reliable evidence. The information provided in the statement aligns with credible sources and can be confirmed as true.
2. Mostly True: When a statement contains a minor inaccuracy or omission, but it is overall consistent with the available evidence, it receives a "Mostly True" rating. While some details may require clarification, the general message of the claim is accurate.
3. Half True: Statements that have elements of truth but are presented in a way that may be misleading or lacking crucial context receive a "Half True" rating. These claims may include partial truths or interpretations that could lead to a misunderstanding.
4. Mostly False: Claims that have some degree of truth but are significantly distorted or taken out of context are labeled as "Mostly False." While elements of accuracy may be present, the overall narrative or implications of the statement are misleading.
5. False: Statements that are demonstrably inaccurate and contradicted by credible evidence receive a "False" rating. These claims are not supported by reliable information and are often intentionally misleading.
6. Pants on Fire: The most severe rating, "Pants on Fire," is reserved for statements that are not only false but also egregiously misleading and designed to deceive. Claims that receive this rating are often based on fabricated information or conspiracy theories.

PolitiFact's six-level rating system provides readers with a nuanced understanding of the accuracy of statements, allowing them to quickly gauge the reliability of information presented in news articles and public discourse. This approach promotes transparency and accountability in public communication by holding individuals and organizations accountable for the accuracy of their claims.

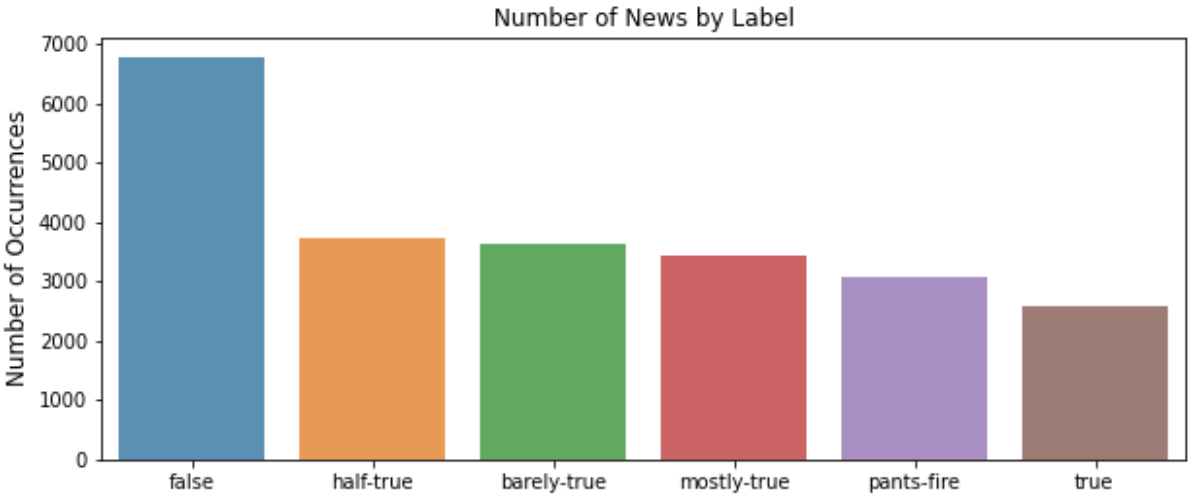


Figure 7-Distribution of News by Labels

In this case, the text and the metadata related to the news was extracted and iterated over all the pages available on the website and for each rating, then the results were combined into a dataframe composed by text, metadata, author, and label.

In conclusion, the dataset for a fake news detection model was meticulously built by harnessing the capabilities of Python's Beautiful Soup library. The utilization of this resource-rich platform not only provided a diverse range of news statements but also ensured that the dataset's content was sourced from a reputable and reliable authority in fact-checking. The final dataset comprising has a total of 23217 rows and 4 columns: Text, Metadata, Source (author) and Label. The distribution of data by occurrences shows that most of the data gathered belongs to the "False" label, with almost 10000 data points. The other labels by count are ordered respectively in the following way: half-true, mostly true, pants on fire, true and half-true.

5.2 Data cleaning and processing

Data cleaning and processing play a pivotal role in constructing a robust and reliable fake news detection model. Raw data gathered from diverse sources can often be riddled with noise, inconsistencies, and extraneous information that might compromise the effectiveness of the subsequent analysis. To mitigate these challenges, a comprehensive data cleaning and processing pipeline is employed, aiming to refine the data into a format that is amenable to accurate analysis.

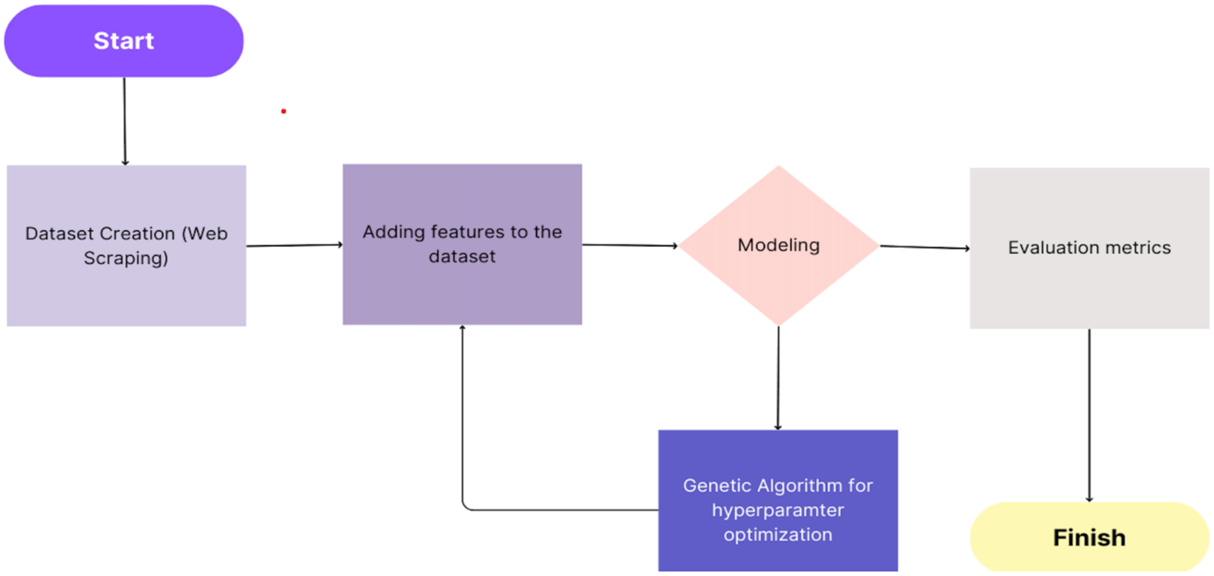

Start
Dataset Creation (Web Scraping)
Adding features to the dataset
Modeling
Evaluation metrics
Genetic Algorithm for hyperparamter optimization
Finish

Figure 8- Flowchart of the proposed model

The process involves preparing text data for analysis or model training by eliminating irrelevant information and normalizing the text thanks to libraries like sklearn and nltk.

One of the initial steps involves the removal of duplicate entries, which helps prevent bias and ensures that each piece of information contributes uniquely to the model's learning process. Handling missing values is another critical aspect, involving techniques such as imputation or elimination to ensure that the dataset is complete and representative.

Standardizing the text format is essential to harmonize the text data so the text converts all characters to lowercase and eliminates any inconsistencies in formatting. Special characters, punctuation marks, and non-essential stopwords are removed to focus on meaningful content while reducing dimensionality. Stopwords are words that are commonly used in a language but generally hold little intrinsic meaning on their own, such as "and", "the," "of," "in," "to," and others. These words are essential for grammatical structure and sentence formation but often don't contribute significantly to the context or semantics of a text. In natural language processing (NLP) and text analysis tasks, stopwords are removed from text data during preprocessing to focus on the more meaningful words that can carry the essence of the content.

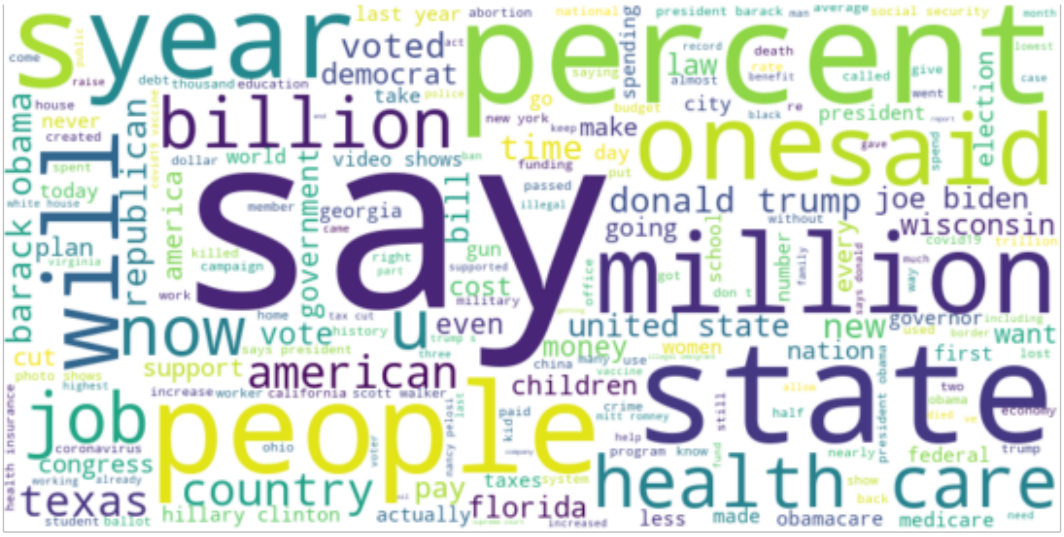


Figure 9- Word Cloud

Furthermore, since this will be handled as a classification task, labels must be turned into a binary variable because the current values reflect the Truth-o-meter ones. For this reason, each label was given a 0 in case of untruthfulness and 1 in case of true news. Labels like 'barely true', 'false', and 'pants-fire' are all mapped to binary value 0, whereas labels like ‘half-true’, ‘true' and 'mostly-true' were given the binary value 1. The composition of data by label (Fig 7) shows that the class containing the largest number of messages corresponds to the “False” (6785 observations), then respectively “Pants on fire” (3068), “Half true” (3723), “Mostly True” (3425 rows), “True” (2585) and “Barely True” (3631).

The data exploration related to the dataset was carried out through the word cloud and n-gram analysis. The former summarizes all the most recurring words in a text, in this case series of sentences; n-grams instead are sequences of text of n words. Even though the dataset does not describe the topic around which news revolve, the word cloud (Fig. 10) suggests that the predominant theme is politics. Indeed, words like “state” or “Donald trump” are among the most frequent ones, the same occurs in the n-gram analysis.

5.3 Feature engineering

Although linguistic characteristics have been successful in categorizing news as true or false, there are no clear standards for choosing the features that are most suited in various situations (Puraivan et al., 2023). A thoughtful selection of linguistic characteristics may considerably aid in the analysis process and improve the results' interpretability. In Puraivan et al. (2023) study, 88 linguistic aspects are gathered in their whole and are carefully categorized into four groups: surface information, part of speech, discursive characteristics, and readability indices.

1. Surface information: numerical data that has been obtained by descriptive statistics, especially summaries and means across various phrases, paragraphs, words, and characters.
2. Part of Speech (POS): data on POS tags, including descriptive metrics like sums, means, minimums, maximums, standard deviations, and medians. The frequency of morphological elements, such as adjectives, adverbs, pronouns, punctuation, and verbs, among others, in the text is also quantified (idf, tf-idf).
3. Discursive Characteristics: computation of frequency for both morphological and discursive categories. For instance, these include words with obnoxious or rude meanings, determiners, demonstratives, personal pronouns, adverbs, articles, prepositions, and negations, as well as functional components like discursive markers and different connectors. Along with linguistic and psychological elements included in the Linguistic Inquiry and Word Count (LIWC) program, the tool also allows for the insertion of lexicons relating to positive and negative emotions.
4. Readability Indices: includes many readability indices that offer a numerical evaluation of the readability of a text.

5.3.1 Surface information

Just like anticipated in Puraivan et al. (2023), surface information provides quantitative data from descriptive statistics on paragraphs, phrases, words, and characters. The data collected contains 9733 verified information, therefore true news and 13484 fake news classified as 0s. As it can be easily observed, the dataset is slightly imbalanced. Handling imbalanced datasets is a critical step in building effective machine learning models, particularly in classification tasks.

There are multiple of alternatives in this case to solve the imbalance: one common approach is resampling, which involves either oversampling the minority class or undersampling the majority class. Oversampling involves duplicating instances from the minority class, increasing its representation in the dataset. Undersampling, on the other hand, involves randomly removing instances from the majority class, balancing the class distribution. However, while resampling can help mitigate class imbalance, it may lead to overfitting or loss of information.

Another technique is generating synthetic data through methods like Synthetic Minority Over-sampling Technique (SMOTE). In addition, evaluation metrics also play a crucial role. Accuracy might not be a suitable metric for imbalanced datasets since a model can achieve high accuracy by merely predicting the majority class. Instead, metrics like precision, recall, F1-score, and area under the ROC curve (AUC-ROC) provide a more comprehensive picture of a model's performance on different classes. In order to guarantee a clear interpretation of the model’s performance, undersampling was the solution chosen (9733 instances for each label).

From the boxplot in Fig. 10, it is possible to observe that the distribution of the length for each news in the dataset is concentrated between 0 and 500, precisely between 0 and 200. The median is around 98 which indicates that half of our sentences are below this length and half are above, furthermore the range of the data, as indicated by the whiskers of the boxplot, extends from about 17 words to 219 words which means that there is high variability in length. Another additional feature used in the model regards how many letters there are in the words of a given text. Here, as expected, the boxplot shows a distribution going from around 3 to 7, where the median lies in more or less the center of the interval (equal to 5). Additionally, the number of words per sentence ranges from around 4 to around 45 so in general sentences are not that short. In general, the dataset seems to be representative of typical news articles where sentence lengths are moderate, and word lengths are reasonably standard.

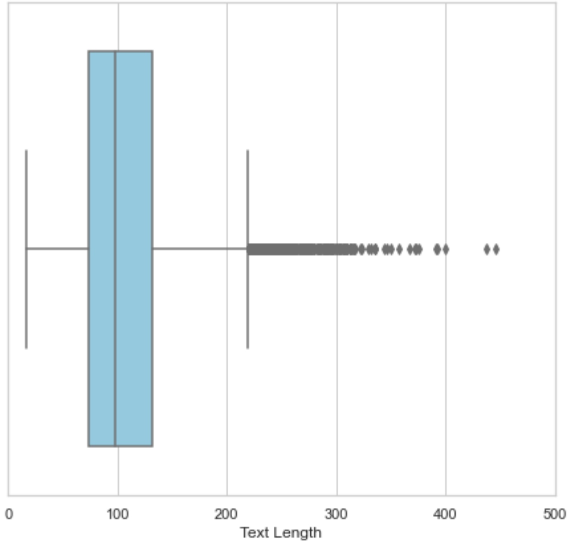


Figure 10- Boxplot of text length

5.3.2 Part of Speech (POS)

Part of Speech (POS) tagging in Natural Language Processing (NLP) is the process of designating each word in a text (corpus) as belonging to a certain part of speech, depending on both its meaning and its context. Identification of a word's grammatical group, such as whether it is a noun, verb, adjective, or adverb, is the main objective of POS tagging. The POS analysis of the dataset was performed by using the nltk package, specifically the “averaged perceptron tagger” function. As expected, the count of adjectives per sentences is mostly between 0 and 2, whereas adverbs are not frequently used: the majority of sentences around 7501 observations do not have adverbs. Among the other characteristics of the text, also pronouns were analyzed, leading to an almost equal results as the adverbs. Same actions were performed for the number of verbs and punctuation.

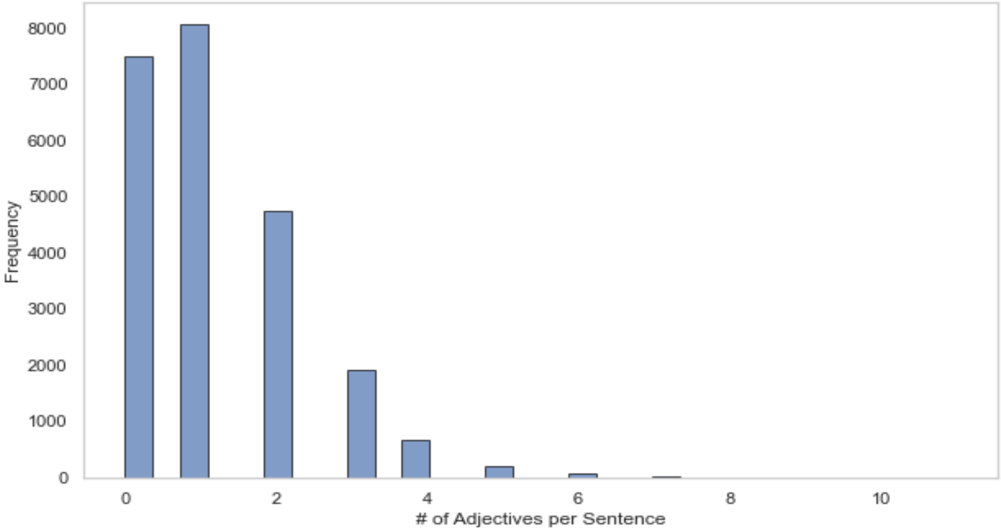


Figure 11- Distribution of adjectives

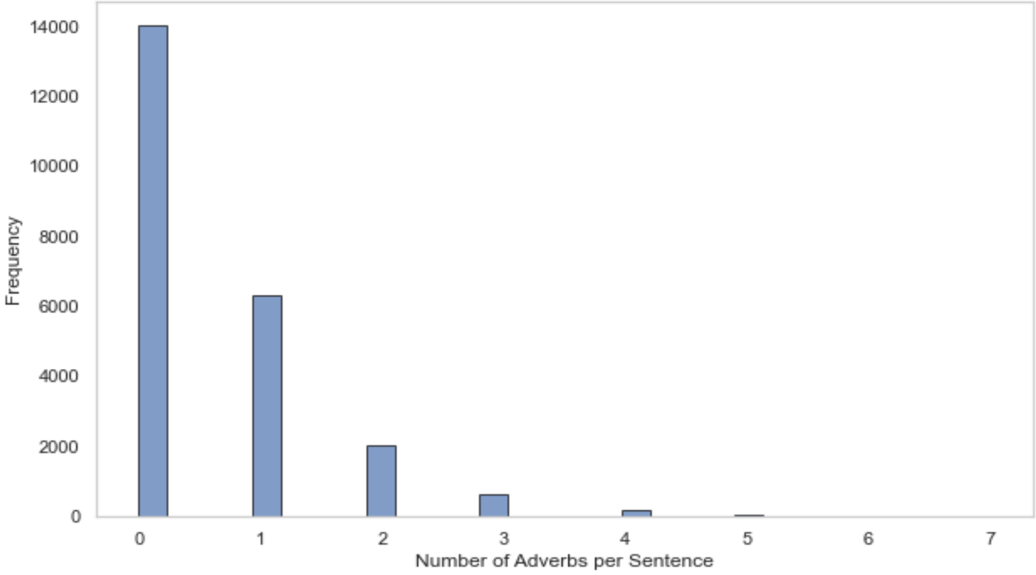


Figure 12- Distribution of adverbs

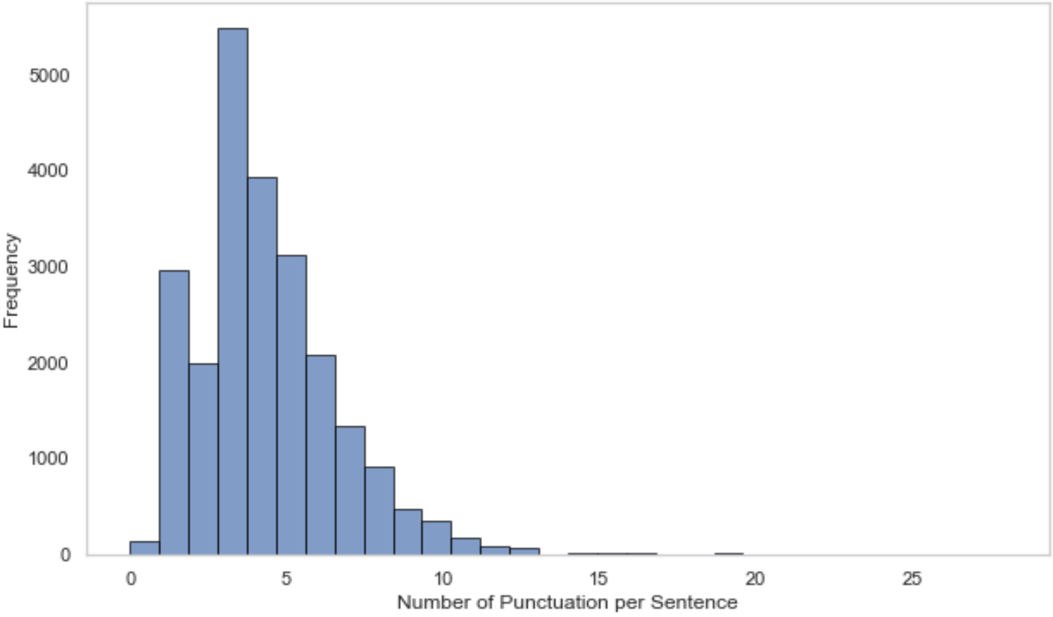


Figure 13- Distribution of Punctuation

5.3.3 Discursive Characteristics- Sentiment Analysis

As Puraivan et al. (2023) anticipates, some characteristics of discourse entail the positive/negative emotions the news is able to evoke.

Sentiment analysis is a computational technique used to determine the emotional tone or sentiment expressed in a piece of text. It involves the use of natural language processing and machine learning methods to classify the sentiment of a text as positive, negative, or neutral. In the context of fake news detection, sentiment analysis plays a crucial role as it provides insights into the emotional content of the news articles. By analyzing the sentiment expressed in a news article, one can assess whether the language used is emotionally charged, sensationalized, or biased, which are common traits of fake news. Sentiment analysis can help in identifying articles that use exaggerated emotions to manipulate readers or to evoke certain reactions. Among the psycholinguistic and semantic-syntactic features for finding fake reviews sentiment features have demonstrated a great potential (Hajek et al., 2023).

The tool used in the modeld devised was TextBlob which is a python library for Natural Language Processing (NLP). TextBlob effectively leverages the Natural Language Toolkit (NLTK) to accomplish its objectives. NLTK is a library that provides convenient access to numerous lexical resources, enabling users to engage in tasks such as categorization and classification. In this context, TextBlob stands out as a user-friendly yet robust tool capable of intricate textual data analysis and operations. Sentiment analysis using lexicon-based techniques depends on the semantic orientation and intensity connected to each word in a phrase (Shah, 2020).This involves using a predetermined lexicon that divides words into negative and positive categories. Typically, a bag-of-words technique is used to represent text information. A pooling procedure consisting in the computation of sentiment for each word summarized into a unique score of the sentence by taking the average value.

In particular, the TextBlob features that were added to the study were two: polarity and subjectivity.

Polarity comprises values ranging from -1 to 1 where -1 stands for a negative sentiment and the opposite in case of 1. Negation words have the effect of flipping the polarity of the statement (Shah, 2020). Semantic labels are used by TextBlob to improve the accuracy of sophisticated analysis. For example, it can distinguish things like emoticons, exclamation points, and emojis, which helps with a more thorough assessment. Subjectivity, on the other hand, ranges from 0 to 1, and it serves as a gauge for how much of the text is made up of personal opinion as opposed to reality (Shah, 2020). Higher subjectivity scores suggest a higher level of subjective information compared to objective data. Intensity is the essential parameter through which subjectivity is computed and it can be defined as the power of one word of modifying the next one (i.e., “very good”).

A special case occurs in the case in which no words of a sentence are contained in the pre-defined text: this leads to 0 for both subjectivity and polarity. As it can be expected, most sentences do not have a clear score in terms of subjectivity and polarity because most words are missing from the established dictionary of the TextBlob package, but the remaining instances demonstrate to have a rather positive sentiment.

5.4 Genetic Search

Genetic search, inspired by the process of natural evolution, offers an innovative approach to hyperparameter optimization in machine learning models. Generally, genetic algorithms are based on Darwin's theories. In his formulation, natural selection sis the "principle by which each slight variation [of a trait], if useful, is preserved." (Darwin, 1859). The notion, while straightforward, holds immense significance: organisms most suitably adapted to their surroundings tend to thrive and propagate, a concept often encapsulated as "survival of the fittest" (Rahman, 2020).

Consequently, individuals best adapted to their environments are more likely to survive and reproduce.

The application of genetic algorithm for fake news detection purposes was also carried out by Okunoye et al. (2022) which uses deep learning models, reaching an accuracy of 74% for fake news and 56,56% for reliable news.

Traditional grid search and random search methods can be computationally expensive and may not always find the optimal set of hyperparameters, especially in high-dimensional spaces. Genetic algorithms, on the other hand, work by initializing a population of potential solutions (individuals), each representing a unique set of hyperparameters. Over successive iterations (generations), these solutions undergo processes analogous to selection, crossover (recombination), and mutation (Russell & Norvig, 1962-. (2010)). The fittest individuals—those that produce the best model performance—are selected to produce offspring for the next generation. Over time, this iterative process of evolution tends to converge towards an optimal or near-optimal set of hyperparameters. In the set of k initial states, forming the population, each state, referred to as an individual, is depicted as a string over a finite alphabet, commonly as a sequence of 0s and 1s. Alternatively, the state might be represented as 8 digits spanning from 1 to 8.

As Russel and Norvig state, this particular GA variant employs a reproduction selection probability that is proportionate to the fitness (1962-. (2010)) score. Two pairs are randomly chosen for reproduction based on probabilities. Notably, a single individual can be chosen twice, while another can be excluded. For each mating pair, a crossover point within the string positions is randomly selected.

Subsequently, offsprings are generated through the crossover of parent strings at the designated crossover point. This means that the newborns contain both strings from parent 1 and parent 2 combined. This example highlights that the crossover operation can yield offspring that diverge significantly from either parent, particularly when the parent states are dissimilar. As previously anticipated, the "fittest" individual survives, and this evaluation is carried out through the definition of a fitness function.

Finally, each position is subject to random mutation with a slight independent probability.

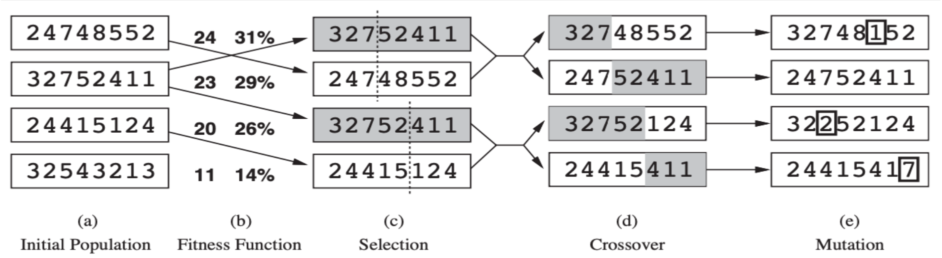
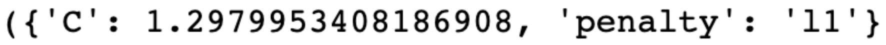


Figure 14- The genetic algorithm represented by digit strings (Russell, S. & Norvig, P. 1962-. (2010))

Genetic algorithms primarily gain from the crossover operation, which introduces the potential for advantageous permutations, and this can be inferred by the image showing each part of the algorithm since the crossover determine the random splitting of the offsprings (Fig., 14).

Due to the high computational power required, the genetic algorithm for hyperparameter tuning was used only on some algorithms: logistic regression, random forest, and LSTM. For the models optimized, the fitness function is represented by the AUC score. Using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) over accuracy as a performance metric can offer several advantages: for instance, in the case of imbalanced datasets, accuracy can be misleading due to the fact that the majority class will have high accuracy and the minority one not. Furthermore, the accuracy has a fixed threshold usually set at 0.5 instead the AUC evaluates the model's performance across all possible thresholds, and this may lead to poor evaluation of false positives and negative rates. In general, using AUC-ROC over accuracy can provide a more robust and informative evaluation of a model's performance.

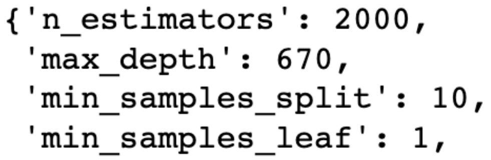
For the logistic regression the hyperparameter optimized were the C (parameter that controls the amount of regularization) and the penalty, parameter added to the coefficients of the regression to lower the magnitude and prevent overfitting. The logistic regression best parameters are:



```
({'C': 1.2979953408186908, 'penalty': 'l1'}
```

corresponding to an AUC of around 0.75 which is considered optimal for a fake news detection model. A C parameter equal to approximately 1.3 suggests a relatively balanced approach to regularization, helping to prevent overfitting without adding too much bias. The 'l1' penalty indicates Lasso regression, which adds an L1 norm of the coefficients to the loss function.

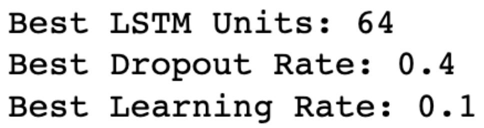
As for the Random Forest, the parameters chosen to be optimized are five: number of trees in the forest, depth of the tree, minimum number of samples required to split an internal node, the minimum number of samples needed to reach a leaf node, bootstrap and the number of features considered for the best split.



```
{'n_estimators': 2000,
 'max_depth': 670,
 'min_samples_split': 10,
 'min_samples_leaf': 1,
```

Figure 15- Genetic optimization results on Random Forest

A choice of 2000 trees is quite high, implying that the model will be computationally intensive. However, this also means that the model has a larger number of weaker learners to make a decision, which generally can increase the model's performance in terms of accuracy and stability Maximum depth equal to 670 means that the trees in the forest can have a substantial depth, allowing the model to learn very detailed patterns in the data. Finally, the maximum features equal to log2 means that the algorithm will consider the logarithm base 2 of the number of features at each split. As for the LSTM, the features optimized were three, the units, dropout rate and learning rate. Genetic optimization reveals that the best LSTM units to use are 64 which means that the network has a good balance of learning capacity and computational efficiency, and it is capable of learning complex patterns in the data without being too computationally expensive. The best dropout rate is equal to 0.4, which helps the model to generalize better and preventing overfitting during training, and finally a learning rate of 0.1 which is relatively high for an algorithm because it means that it learns quickly but it has the risk of overshooting the optimal solution.



```
Best LSTM Units: 64
Best Dropout Rate: 0.4
Best Learning Rate: 0.1
```

Figure 16- Genetic optimization Results on Random Forest

5.5 Algorithms

The algorithms applied to this research can be divided into three macro categories: traditional algorithms, neural networks, and NLP. According to the literature review in chapter 4 the best approach to obtain high results in fake news detection models are NLP-based. However, every algorithm must be tailored to the dataset and resources available. Most of the code has been executed on a macOS Ventura with an Intel Core i5 processor 1.6 GHz and, due to the high computational power, the most requiring steps (like Bert models) were executed on Colab Pro+ using an A100 NVIDIA GPU. Python 3 was used to set the configurations of all the models and multiple libraries were used like Keras, PyTorch, Transformers and NLTK. Considering the small size of the dataset, one of the first assumption to be made is that simpler models should perform better than more complex ones. As a consequence, overfitting can cause some problems in this case, and it is fundamental utilizing regularization terms and dropout layers to minimize the issue.

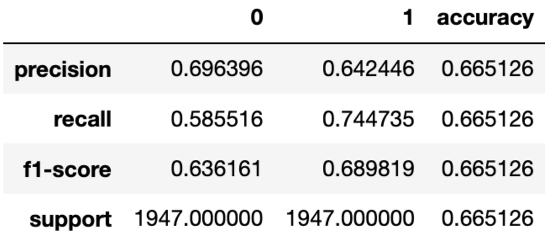
5.5.1 Traditional models

The traditional models applied in this research are Logistic Regression, Random Forest, Decision Tree, GaussianNB, and SVM. These widely used algorithms each bring their own strengths and characteristics to the table when addressing the specific challenges of the given problem domain. When dealing with fake news detection, each of these frequently utilized algorithms brings its advantages and traits to the table.

Logistic regression is the simplest model employed compared to the other mentioned since it is an extension of the linear regression model. This will be therefore used as baseline for the other models since it is not very suitable for complex data and does not capture the interactions among features. The accuracy for this model is around 67%. However, by looking closer at the classification report, it is important to highlight that the precision for class 0 is about 68% so when predicting false news, the model is pretty accurate. The recall remains at 62% but by looking at class 1 metrics precision for class 1 is about 65% while the score for the recall is even higher, approximately 71%. Out of 10 the actual class 0 instances, 7 times the model is correct. Overall, the model seems to perform reasonably well, with balanced precision and recall for both classes. However, the F1-scores indicate a slight imbalance in favor of class 1, as can be also noticed by the statistics.

Instead, a decision tree depicts data in a tree-like structure, with each leaf node (or split) denoting the anticipated class or value and each internal node (or split) denoting a choice based on a characteristic. The rules that direct the decision-making process are defined by the path from the root node to the leaf node. Usually, one of the main drawbacks caused by this model is overfitting because it ends up generalizing too much data. Precision for the Decision Tree is 59.6% which means that more than half the news articles predicted as class '0' were actually class '0', it can be considered a moderate level of precision. The recall suggests that about 61.5% of the actual class '0' articles were correctly identified by the model. This implies that the model is slightly better at capturing the actual '0' class instances at the cost of including more false positives. For class 1, about 60.2% of the instances predicted as class '1' were indeed class '1' (precision) and recall is slightly lower compared to the one of class 0.

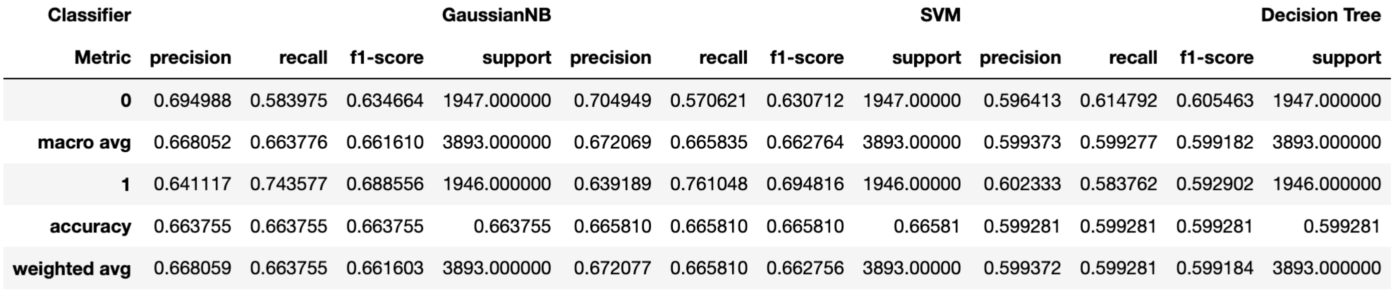
The evolution of the Decision Tree is the Random Forest which is an ensemble learning approach that creates several decision trees and combines their forecasts to increase the model's overall performance and resilience. Contrary to the decision tree technique, the ensemble approach reduces overfitting. Furthermore, Random Forest implicitly performs feature selection and generates uncorrelated decision trees, and this greatly influences the performance of the models. However, being a more complex algorithm, it also gets more difficult to interpret. The overall accuracy of the Random Forest model is approximately 67%, indicating the ratio of correct predictions to the total predictions made. The predictions for class 0 reveal that precision is equal to 70% and recall 59%: this means that the actual fake news is well-recognized. Class 1 metrics are instead slightly: precision around 64% and recall goes until 74%.



	0	1	accuracy
precision	0.696396	0.642446	0.665126
recall	0.585516	0.744735	0.665126
f1-score	0.636161	0.689819	0.665126
support	1947.000000	1947.000000	0.665126

Figure 17-Random Forest Classification

Another algorithm implemented was Gaussian Naïve Bayes which is given by Bayes' theorem. Just like the Logistic regression, this is a very simple model and as such, it may entail different biases in the predictions. It's particularly well-suited for small to moderately-sized datasets and serves as a good starting point for classification tasks, but it's important to carefully evaluate its performance against more complex models when appropriate. Overall, Gaussian Naïve Bayes performs similarly to the logistic regression and does not add much value to the previous model. Finally, among the traditional models, there is also the Support Vector Machine, a versatile and powerful machine learning algorithm used for both classification and regression tasks. SVM seeks to find the hyperplane that best separates classes in the feature space. This model is usually implemented for high-dimensional data, and it is very robust to outliers, and it also handles imbalanced data. Of course, overfitting is always a limitation of modeling, and it has to be taken into account. Indeed. The performance of the model is superior to the models non-optimized. In conclusion, by looking at the totality of metrics SVM and Random Forest are the ones reaching better results and demonstrating superior performance in various aspects of the classification tasks. Their robustness and ability to handle high-dimensional data make them standout choices for this particular analysis. Moreover, their



Classifier	**GaussianNB**				**SVM**				**Decision Tree**			
Metric	**precision**	**recall**	**f1-score**	**support**	**precision**	**recall**	**f1-score**	**support**	**precision**	**recall**	**f1-score**	**support**
0	0.694988	0.583975	0.634664	1947.000000	0.704949	0.570621	0.630712	1947.00000	0.596413	0.614792	0.605463	1947.000000
macro avg	0.668052	0.663776	0.661610	3893.000000	0.672069	0.665835	0.662764	3893.00000	0.599373	0.599277	0.599182	3893.000000
1	0.641117	0.743577	0.688556	1946.000000	0.639189	0.761048	0.694816	1946.00000	0.602333	0.583762	0.592902	1946.000000
accuracy	0.663755	0.663755	0.663755	0.663755	0.665810	0.665810	0.665810	0.66581	0.599281	0.599281	0.599281	0.599281
weighted avg	0.668059	0.663755	0.661603	3893.000000	0.672077	0.665810	0.662756	3893.00000	0.599372	0.599281	0.599184	3893.000000

Figure 18- Non-optimized Algorithms Results

versatility in tuning parameters allows for more precise model optimization, potentially leading to even more accurate and reliable predictions in future studies.

5.5.2 Neural Networks

Given the size constraints of the dataset, deep learning approaches were severely limited. As the adjective indicates deep learning is characterized by neural networks with many layers, hence "deep". Deep learning models have the innate ability to recognize deep patterns and structures in data because of their complex designs with many hidden layers. This skill comes at the expense of needing a lot of data to train efficiently, though. These models are prone to overfitting in circumstances with little data, whereby they effectively recall the training data and fall short in their ability to generalize to new data. This problem is resolved by larger datasets, which provide a broad variety of instances that help the model learn more reliable and complex characteristics. Therefore, having only around 23000 data, the models were built with few layers and careful attention was put on the overfitting problem.

The first model implemented is a hybrid LSTM because both text data and numerical ones have been put as inputs into the LSTM. So, the first input layer of the LSTM has been configured for the text data with 64 units as the optimization step suggested. Moreover, for the additional features, a feed-forward neural network is implemented with the same number of units of the text data. The final output layer contains both text and features and, by applying a sigmoid activation function, the model returns probabilities. The probabilities define the percentage of belonging to a certain class. In this case the higher the probability, the higher the possibility of having true news. The accuracy of the model is equal to approximately 62%. By looking at overfitting it can be noticed that validation loss always keeps steady around the same values, and this does not necessarily indicate overfitting. Overfitting is typically characterized by a continuous decrease in training loss alongside an increase in validation loss, showing that the model is getting better at fitting to the training data but worse at generalizing to new, unseen data. Also, the gap between the two losses is not very high and thus, it can be concluded that the performance limitation purely regards the size of the dataset.

For this reason, the early stopping parameter was added to the model to monitor overfitting. The tailored function for the early stopping monitored the delta between train and validation loss. The model performed all the 100 epochs set at the beginning without stopping which leads to the conclusion that overfitting is not present. over 100 epochs without stopping. Anyway, it is also worth looking at the other metrics: on the validation set the recall reaches 70% and precision 62%.

The performance of the model was limited due to the size constraints. Therefore, as subsequent steps, the model complexity was lowered.

Indeed, when a simple neural network is applied to the model, the results vastly improve, bringing the validation accuracy to 67% but with excellent metrics on recall and f1 score (respectively 82% and 71%). In this case, no overfitting is present. Since validation and training loss are minimized are close to each other. Consequently, when trying to apply a simple LSTM on textual data the result leads to 65% accuracy, showing that small-medium sized datasets need less model complexity.

5.5.3 NLP: language modeling

5.5.3.1 BERT Introduction: LM architecture

Bert stands for Bidirectional Encoder Representations from Transformers which means that it has the same structure as large language models, and it is based on the attention mechanism, or better the multi-head attention described in chapter 3. The key revolution of this model is the training of a bidirectional transformer (Horev, 2018). Before BERT innovation, the other algorithms looked at a text sequence either from left to right or the opposite. According to Horev (2018), bidirectional training can indeed increase the awareness of the model on the context. As the name suggests, BERT includes two separate mechanisms: an encoder that gets the input and a decoder generating the prediction. One of the main key advantages is the "non-directionality" since the attention mechanism is able to understand the context by enhancing the words to focus more on. The input consists of a series of tokens, initially embedded into vectors, and subsequently processed in a neural network. The result is a collection of H-sized vectors, each of which aligns with a token from the input that has the same index.

Devlin et al. (2019) divide BERT framework into two big macro categories: pre-training and fine-tuning.

Language modeling has always been based on a predictive approach, which means that the majority of the models before BERT invention aimed at forecasting the subsequent word in a sentence and were trained either from left to right or from right to left. For this reason, BERT adopts two training strategies: Masked LM and Next-sequence prediction (Horev, 2018).

Before inputting word sequences into BERT, around 15% of the words in the sentences are substituted by some mask tokens that obscure the real ones. Based on the context extracted from the other words, non-masked, the model then makes an effort to forecast the original value of the masked words. It is important to specify that in order to avoid a mismatch between pre-training and fine-tuning, only 15% of the positions are predicted and out of this amount not all tokens are replaced with MASK token (Devlin et al.2019):

1. 80% of the time the [MASK] token takes the place of the i-th token.
2. In 10% of cases, it's replaced with a random token.
3. Another 10% of the cases, it does not change.

Basically, what happens consists of building a classification layer upon the encoder output. The output vectors of the encoder are multiplied by the embedding matrix and form some vocabulary dimensions. The masked word is determined by computing the probability for each word in the sentence. The peculiarity is that only masked words are taken into consideration for prediction. Consequently, the model optimizes the convergence point.

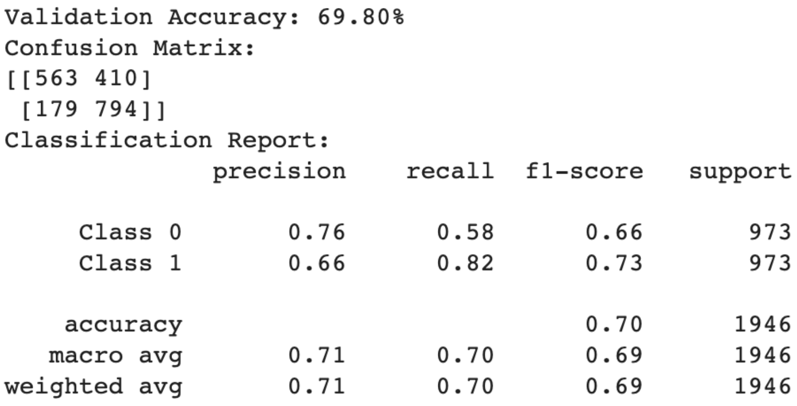
The second strategy is called Next-Sentence Prediction (NSP) and consists in learning what sentence goes after the first. Indeed, in this type of training, 50% of the inputs are the original pairs of inputs and the other 50% of the two sentences are randomly chosen. The base belief here is that two sentences randomly concatenated do not make sense. In order to signal the model where one sentence starts and ends, some tokens are added to the sequence: CLS token at the beginning and SEP token at the end. Then two embeddings are performed on the sentence: sentence embeddings and positional ones. At this point, the transformer is fed with the input and the output of the CLS token is returned by a classification layer and the next sequence is just a matter of probability (Horev, 2018). In the end, the model trains both using the Masked LM and Next Sequence Prediction function aiming at minimizing the loss of the two functions. Horev (2018) underlines the importance of having enough training data for Language modeling to reach a good accuracy.

Fine-tuning BERT, or any other pre-trained model, refers to the process of taking a pre-trained model and adjusting or "tuning" it slightly with additional training to make it more suited for a specific task. Thanks to the self-attention mechanism it is possible to optimize the bidirectional attention mechanism. Fine-tuning Bert, according to Devlin et al. (2019) is relatively easy since it is enough to input the data on which the task has to be executed and adjust each parameter.

5.5.3.2 BERT Results

In the process of curbing the proliferation of fake news, the application of advanced models like BERT (Bidirectional Encoder Representations from Transformers) can play a pivotal role.

Initially, undersampling is employed to balance the dataset, mitigating the bias towards the majority class which usually encompasses genuine news articles, thereby fostering a more balanced learning environment. Following this step, the BERT tokenizer comes into play, which is adept at understanding the shades of language in textual data, transforming sentences into a format that's amenable to model training. This tokenizer not only segments text into words but also considers the contextual relationship between words, a feature that is quintessential in identifying the subtle markers of misinformation. Subsequently, preprocessing is executed to cleanse and structure the data efficiently, facilitating a smoother and more focused model training process. By combining the robustness of BERT's linguistic understanding with meticulous data preprocessing, the model can potentially excel in discerning fake news with high precision and reliability. This approach essentially amalgamates data engineering techniques with deep learning methodologies to foster a comprehensive strategy against the dissemination of false information. Given the model’s robustness and computational requirements, on such a small dataset with social media messages, we would expect to have a moderate ability to discern truth from false. In order to prepare the dataset for BERT model, attention masks, which are binary masks indicating the positions of the tokens that should be attended to (as opposed to padding tokens), are generated and converted, along with the input ids, into tensors. Thanks to the use of the dataloader the model will train and validate in batches, loading parallelly data and accelerating the training process, but most importantly, optimizing the usage of memory and computational power.

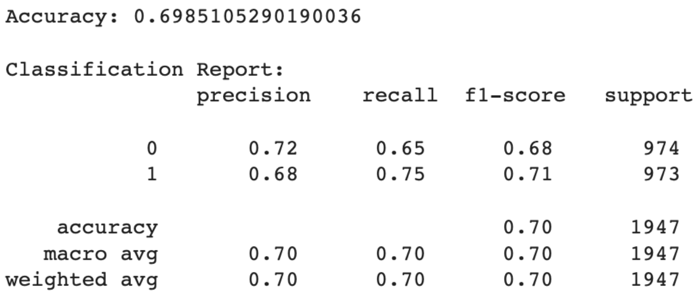


```
Validation Accuracy: 69.80%
Confusion Matrix:
[[563 410]
 [179 794]]
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.76      0.58      0.66       973
     Class 1       0.66      0.82      0.73       973

    accuracy                           0.70      1946
   macro avg       0.71      0.70      0.69      1946
weighted avg       0.71      0.70      0.69      1946
```

Figure 19-BERT Results

The model correctly identified the class (whether positive or negative) of about 69.8% of the total instances. This metric gives a general idea of the model's performance, but it should be considered alongside other metrics, especially when the dataset is imbalanced. Of all the instances that the model identified as the positive class, about 66% were actually of the positive class. It indicates that the model has a relatively moderate rate of false-positive errors. The model correctly identified about 82% of all actual positive instances. This high recall indicates that the model is proficient at detecting the positive class but at the expense of a higher number of false positives (as indicated by the lower precision).

Overall, the model seems to have a decent performance with a good recall rate, indicating its capability to identify the positive class effectively. However, the precision is somewhat lower, indicating a higher rate of false-positive errors. Depending on the specific context and the relative costs of false positives and false negatives, further tuning might be necessary to improve the model's precision without significantly sacrificing recall. NLP confirms in this case its superiority in terms of news classification. By further simplifying the model and applying a DistilBERT results are even more balanced. The "distil" in its name refers to the process of distillation, where the model is trained to imitate the behavior of a larger, more complex model (in this case, BERT). Also in this case, accuracy revolves around 70%. The relatively high recall score indicates that the model is proficient at detecting positive instances, which might be particularly important in the context of fake news detection. The F1 score of 70%, providing a single score for precision and recall, reflects the good balance between the two measures.



```
Accuracy: 0.6985105290190036

Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.65      0.68       974
           1       0.68      0.75      0.71       973

    accuracy                           0.70      1947
   macro avg       0.70      0.70      0.70      1947
weighted avg       0.70      0.70      0.70      1947
```

Figure 20-DistilBERT Results

5.4 Evaluation Metrics

Given the size of the dataset, the evaluation metrics used for these models cannot be limited to accuracy. Getting more insights into the classification of both classes is essential to guarantee the consistency of the model. Before delving into the metrics, it is useful to determine the confusion matrix, a matrix that compares the actual value of the dataset and the predicted values. It is an n*n matrix where n is the number of classes. The possible outcomes in this matrix are (Goyal, 2021):

1. True Positives (TP): number of instances actually positive that are predicted positive. In this case, considering positive a true news, the number of news actually real predicted correctly.
2. True Negatives (TN): number of instances actually negative that are predicted negatively. This is the opposite case to the TP, when a piece of fake news is correctly identified as such by the model.
3. False Positives (FP): also called type I error, defines the number of negative instances that are predicted positive. This would occur when information is fake but considered true.
4. False Negatives (FN): known as type II error represents the number of instances actually positive that are predicted negative. For instance, true news is classified as fake.

From the confusion matrix, it is possible to extract 4 evaluation formulas. The first metric used is accuracy which identifies the ratio between the number of correct predictions and the total number of predictions made. For this reason, it is a suboptimal solution when considering an unbalanced dataset. In order to have a bigger and deeper overview of the dataset, it is necessary to broaden the analysis.

$$Accuracy = \frac{\#\ of\ correct\ predictions\ (TP + TN)}{total\ number\ of\ predictions\ made\ (TP + TN + FP + FN)}$$

The second metric that will be used for the analysis of these models is precision which is defined as the ratio between TP divided by all the instances predicted as positives: the number of real positives out of all the predicted positives.

$$Precision = \frac{TP}{TP + FP}$$

Then, the third is recall, a measure that defines the TP out of the real positives. If the result is high, it means that the model is able to detect positive samples.

$$Recall = \frac{TP}{TP + FN}$$

From the formulas, it is visible the dependency that there is between precision and both positive and negative samples. Instead, recall is dependent only on the positive samples.

Another of the metrics mostly used is F1-score. The latter is a combination of both precision and recall, the higher the score the better. This score is high only both variables are high.

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Other metrics usually implemented in the analysis of a dataset are the ROC (receiver operating characteristic curve) curve and AUC (area under the ROC curve). The ROC is a representation of recall (also known as True Positive Rate) and False Positive Rate. The AUC measures the degree to which the model is capable of separating the two classes while ROC is its probability curve (Narkhede, 2018). In conclusion, the higher the AUC, the better the model is at distinguishing classes.

5.5 Results & Further Research

In the burgeoning field of machine learning, the work of Kondamudi et al. (2023) stands as a comprehensive map, tracing the intricate networks of both supervised and unsupervised learning methodologies. Amidst the proliferating challenges of online hoaxes, frauds, and misinformation, the study shines a spotlight on several potent supervised machine learning techniques, including Logistic Regression, Random Forest, K-Nearest Neighbors, Decision Trees, and Support Vector Machines (SVM). These techniques have been historically instrumental in discerning and mitigating misleading information online, thus serving as powerful tools in the ongoing battle against digital misinformation. In light of these studies, traditional models, with their unique strengths and characteristics, should not be underrated, especially in the case of small-medium-sized datasets. Logistic Regression, despite its simplicity and limitations in handling complex data, serves as a reliable baseline, exhibiting considerable accuracy, particularly in identifying false news. SVM and Random Forest thanks to their robust nature, coupled with their ability to adeptly handle high-dimensional data emerge as frontrunners, showcasing superior performance in various facets of classification tasks. For instance, Rubin et al. (2017) obtained an 82% on a satirical dataset using SVM. If compared with the results of a proposed study using social media news, satirical news is much more "explicit" in terms of writing patterns and sentiment. Furthermore, simple neural networks have shown great potential in content-based fake news detection. Indeed, the hybrid LSTM model, meticulously balances text, and numerical data inputs to avoid overfitting, as evinced by the stable validation loss and the non-significant gap between training and validation loss. Despite these precautions, the study conceded to the constraints imposed by the limited dataset, manifesting in a relatively modest accuracy of approximately 62%. Consequently, the study ventured into simplifying the model complexity, resulting in improved validation accuracies and metrics, thereby underscoring the efficacy of less complex models in handling small to medium-sized datasets.

Overall, while the literature presents an optimistic panorama of the advancements and potentials of deep learning mechanisms in fake news detection, underscored by substantial accuracies and the adept handling of big datasets, this study underlines a prudent approach when working with limited datasets. It propounds a nuanced perspective that sometimes, less complex models might offer a more viable and effective solution, thereby adding a rich dimension to the ongoing dialogue on the optimization of fake news detection strategies. This narrative, however, takes a fascinating turn when considering the remarkable strides achieved through the utilization of language modeling. As expected, despite the intricate architecture inherent to BERT, the attention mechanism confirms its ability in understanding context better than other methods. The diverse variants of BERT pre-trained model allow you to find the most consistent model for a specific purpose and dataset. Although suffering on some metrics like precision, the overall performance, mirrored in a balanced F1 score of 70%, hints at the model's proficiency in false news detection.

The role of good data quality cannot be overstated in the effort to combat fake news. Simply put, the better the data, the more effectively and accurately we can identify and prevent the spread of misinformation. Some limitations of the study are computational requirements and lack of multi-modal information. It's crucial to note that fake news doesn't just exist in written form; it can be found in images and videos as well. That's why future studies should focus more on multi-modal approaches. By embracing a more comprehensive approach, researchers can develop smarter and more robust tools to fight against fake news, making our information landscape safer and more reliable. Fact-checking organizations should be supported more by social media platforms to guarantee the development of more efficient methodologies and real-time solutions. In this case, the size of the dataset stands as a notable constraint. A limited dataset may not offer a comprehensive representation of the vast and varied landscape of information circulating in the digital realm. It could restrict the model's ability to learn and adapt to the complex patterns and nuances that characterize fake news narratives. In essence, a substantial dataset would aid in avoiding overfitting, where the model becomes too adapted to the training data, failing to generalize well to new, unseen data. Thus, expanding the dataset could be a vital step in bolstering the robustness and reliability of the fake news detection system.

6. Conclusions

In the digital era, the velocity of information dissemination has amplified significantly. The spread of fake news seems to outpace that of true information, a trend documented extensively in the research. These fake rumors often exhibit higher levels of "novelty" (Vosoughi et al., 2018) and induce stronger emotional responses, attributes that enhance their attractiveness and shareability. Moreover, the propensity to share misinformation is also influenced by individuals' habits and their trust in online platforms, with those demonstrating higher trust levels being less critical of the authenticity of the information they encounter (Talwar et al., 2019).

Furthermore, the contemporary media landscape's significant transformation favors the diffusion of unreliable information. Recalling Allcot & Gentzkow's research published in 2017, the critical factors creating this environment are the lowered entry barriers to the media industry, the rise of social media platforms, diminishing trust in traditional media, and escalating political polarization, which foster a fertile ground for misinformation to thrive.

In the midst of this evolving narrative, large language models (LLMs) have emerged as both a boon and a threat. While they offer scalability and ease of deployment, their misuse, incentivized by these very attributes, cannot be overlooked. The ease of deployment is what makes it a powerful tool as well as a damaging one: hallucination, biased content, not real-time, disinformation, and inexplicability are five of its main drawbacks (Sohail et al., 2023). AI hallucination is the process by which an AI model generates inaccurate information and explains it as facts; misinformation can be considered part of the hallucination issue. Sohail et al. (2023) talk about the democratization of knowledge diffusion.

Furthermore, the susceptibility of Open-domain Question Answering (ODQA) systems to misinformation pollution, as illustrated by Pan et al. (2023), is a pressing concern, particularly in the context of emergent topics such as COVID-19, where misinformation can potentially have significant consequences. The study accentuates the vulnerability of these systems to manipulation, especially when faced with questions that lack substantial backing or about topics that are relatively new, and hence lack depth of information. As we navigate this complex and dynamic landscape, the urgency to develop robust mechanisms to combat misinformation cannot be understated. As echoed by various researchers, the scalable solution to combat misinformation effectively lies in harnessing the capabilities of AI itself, which at the same time also serves as a medium of fake news diffusion (Chakraborty et al., 2023, Ruffo et al., 2023).

The model proposed in this research leveraged the help of language modeling for fake news detection on social media data, whose structure is totally different from scientific and satiric articles. Indeed, the difficulty lies in developing a model for brief and informal textual data (Singhal et al., 2019). In this case, data is not high-quality by default. Therefore, as a future direction, the inclusion augmenting the dataset stands as a critical step towards enhancing the robustness and reliability of fake news detection systems, enabling them to better generalize to new, unseen data and avoid the pitfalls of overfitting. This is also stressed in the basics of the transformers model. As explained in section 3, scaling laws delineate the relationship between the resources invested in the model and its performance. As the model scales, it is able to understand more complex patterns and improve its performance. In the presence of two factors parameter count (N) and model dimension (D), increasing only one of them does not increase performance, both must change (Kaplan et al., 2020). Scaling laws can be considered as an actual optimization technique in language modeling. As a consequence, it is important to keep in mind that transformer-based models lead to incredible results only in case model size and parameters count are high. One of the considerations in figure 6 is that for a smaller D (denoting dataset size), performance stops increasing as the model gets more complex and starts overfitting (Kaplan et al., 2020).

The scaling laws property explains exactly the performance limitations of the proposed solution. Two strong limitations of the study are dataset size and computational power. The dataset size is indeed a very small dataset for transformer-based models like BERT and this is why there is no incredible increase in model performance. Due to data quality concerns, data was collected by one source and was not amplified because increasing other sources may have brought more noise to the dataset also in terms of news length and type. This could have hindered the model's ability to learn data patterns. Moreover, diverse data implies higher computational power. Therefore, due to the computational limitations of the device used for the research and the data availability of the source, the dataset was kept small.

As a result, considering all the variables in the research, the most reliable evaluation metrics to consider are precision, recall and AUC that specifically address class classification. Even though language modeling outperforms all the other models, thanks to the smallness of the dataset also traditional models reach good results. Although it encounters challenges in some metrics like precision, Transformer-based model performance reflected in a balanced F1 score of 70% alludes to the model's competence in false news detection. In particular, the hypothesis about the scaling law is further proved by the fact that by reducing model complexity, thus applying DistilBERT the overall performance of the model increases. The same occurs for the neural networks where a simple neural network scores higher than the RNNs. An additional optimization step for future studies could be also feature selection. Indeed being able to gather information on the writer for instance can help in finding new features to add to the model.

In conclusion, AI is the only element that can "save" users from AI itself.

Moreover, the synergy between AI capabilities and human expertise encapsulates a promising strategy for enhancing the veracity and efficacy of fact-checking initiatives. This era calls for a united front, combining technological prowess with educational efforts to cultivate a discerning and informed digital population.

Bibliography

(n.d.). Retrieved from Politifact: https://www.politifact.com/

Acemoglu, D., Ozdagar, A. & ParandehGheibi, A. (2010, February 1). Spread of (mis)information in social networks. *Games and Economic Behavior, 70*, pp. 194-227.

Ahmad, I., Yousaf, M., Yousaf, S. & Ahmad, M.O.. (2020). Fake News Detection Using Machine Learning Ensemble Methods . *Complexity*, 1-11.

Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting fake news using machine learning: A systematic literature review. *arXiv preprint*.

Allcott, H. & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives—Volume 31, Number 2*, pp. 211-236.

Ansar, W. & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights journal homepage: www.elsevier.com/locate/jjimei Combating the menace: A survey on characterization and detection of fake news from a data science perspective Wazib Ansar a,a,∗, Saptarsi Goswam*.

Athira, A. B., Kumar, S. M., & Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence, 122*.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural Machine Translation by jointly learning to align and translate. *arXiv preprint arXiv:1409*.

Benevenuto, F., Magno, G., Rodigues, T., & Almeida, V. (2010). Detecting spammers on Twitter. *4th International AAAI Conference on Weblogs and Social Media (ICWSM), 6*.

Bodaghi, A. & Oliveira, J. (2020). The characteristics of rumor spreaders on Twitter: A quantitative analysis on real data. *Computer Communications, 160*, 674-687.

Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P. & García-Cumbreras, M.A. (2021). Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications. Volume 169*.

Bradshaw, S., Howard, P., Kollanyi, B., & Neudert, L. (2019). Sourcing and Automation of Political News and Information over Social Media in the United States. *Political Communication. 37*, 1-21.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Capuano, N., Fenza, G., Loia, V. & Nota, F.D. (2023). Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review. *Neurocomputing 530*, 91-103.

Ceylan, G., Anderson, I. & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences of the United States of America. 120.* .

Chen, S., Xiao, L. & Kumar, A. (2023). Spread of misinformation on social media: What contributes to it and how to combat it . *Computers in Human Behavior, 141*.

Choudhary, A. & Arora, A. (2021). Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications 169*.

Chu, Z., Gianvecchio, S., Wang, H. & Jajodia, S. (2010). Who is tweeting on Twitter: human, bot, or cyborg? *Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC '10). Association for Computing Machinery, New York, NY, USA*, 21-30.

Chui, M., Hall, B., Mayhew, H., Singla, A. and Sukharevsky, A. (2022, December 6). *The state of AI in 2022—and a half decade in review.* Retrieved from QuantumBlack AI by McKinsey: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

Clayton, K. & Blair, S., Busam, J., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A., Wolff, A., Zhou, A. & Nyhan, B. (2020). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior. 42.*

Cresci, S. (2020). A Decade of Social Bot Detection. *Commun. ACM 1.*

Darwin, C. (1859). On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life.

Del Vicario, M., Bessi,A., Zollo, F., Petroni, F., Scala, A., Caldarellia, G., Stanley, H.E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proc Natl Acad Sci U S A.*, 554-559.

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific reports, 7(1)*, 1-12.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Douglas Heaven, W. (2022, November 18). *Why Meta's latest large language model survived only three days online* . Retrieved from MIT Technology Review: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

Entman, R. (1993). Framing: Toward Clarification of a fractured paradigm. *Journal of Communication, 43*, 51-58.

Erokhin, D & Komendantova, N. (2023). The role of bots in spreading conspiracies: Case study of discourse about earthquakes on Twitter . *International Journal of Disaster Risk Reduction 92*.

Ferrara, E. & Yang, Z. (2015). Measuring Emotional Contagion in Social Media . *PLoS ONE 10.*

Garg, S. (2020). A Study on the Structure of Neural Networks and theMathematics behind Backpropagation.

Geschke, D., Lorenz, J. & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology, 58*, pp. 129–149.

Goyal, S. (2021, July 20). *Evaluation Metrics for Classification Models.* Retrieved from Medium: https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69

Granik, M. & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON),*, 900-903.

Gravel, J., D'Amours-Gravel, M. & Osmanlliu, E. (2023). Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health, vol 1, issue 3*, 226-234.

Gupta, A., Li, H., Farnoush, A. & Jiang, W. (2022). Understanding patterns of COVID infodemic: A systematic and pragmatic approach to curb fake news. *Journal of Business Research 140* , 670-683.

Hajek, P., Hikkerova, L. & Sahut, J.M. (2023). Fake review detection in e-Commerce platforms using aspect-based sentiment analysis . *Journal of Business Research 167*.

Hegselmann, R., Krause, U., & Riehl, J. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation, 5(3)*, 1-24.

Horev, R. (2018, November 10). *BERT Explained: State of the art language model for NLP.* Retrieved from Towards Data Science: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Hsu, T. & Thompson, S.A. (2023, February 8). Disinformation Researchers Raise Alarms About A.I. Chatbots. *The New York Times*.

IBM. (n.d.). *What is natural language processing (NLP)?* Retrieved from https://www.ibm.com/topics/natural-language-processing

IBM. (n.d.). *What is natural language processing (NLP)?* Retrieved from IBM: https://www.ibm.com/topics/natural-language-processing

IBM,. (n.d.). *What is a neural network?* Retrieved from IBM: https://www.ibm.com/topics/neural-networks

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint*.

Kemp, S. (2022, October 20). *DIGITAL 2022: OCTOBER GLOBAL STATSHOT REPORT.* Retrieved from DataReportal: https://datareportal.com/reports/digital-2022-october-global-statshot

Khan, J.Y., Khondaker, M.T.I., Afroz, S., Uddin, G. & Iqbal,A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications, 4*.

Kondamudi, M.R., Sahoo, S.R., Chouhan, L. & Yadav, N. (2023). A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University – Computer and Information Sciences 35*.

Kortschak, H. (2020, November 16). *Attention and Transformer Models* . Retrieved from Towards Data Science: https://towardsdatascience.com/attention-and-transformer-models-fe667f958378

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N. & Akbar, M. (2020). Fake news detection using deep learning models: a novel approach. *Transactions on Emerging Telecommunications Technologies 31(2)*.

Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M, Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J. & Zittrain, J.L. (2018). The science of fake news. *Science.359*, 1094-1096.

Lim, W.M., Gunasekara,A., Pallant, J.L., Pallant, J.I. & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarok ¨ or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education, 21.*

Luceri, L., Cardoso, S. & Giordano, S. (2021). Down the bot hole: Actionable insights from a one-year analysis of bot activity on Twitter. *First Monday 26 (3).*

Möller, J., Trilling, D., Helberger, N. & van Es, B. (2018). Do not blame it on the algorithm: an empirical assessment ofmultiple recommender systems and their impact on contentdiversity. *INFORMATION, COMMUNICATION & SOCIETY 21 (7)*, 959-977.

Ma, J., Gao, W. & Wong, K.F. (2017). Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 708-717.

Marr, B. (2023, May 19). *A Short History Of ChatGPT: How We Got To Where We Are Today.* Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/?sh=74796b1e674f

Mazza, M., Bessi, A., Larrere, R., & Volkovich, Y. (2019). Investigating the difference between trolls, social bots, and humans on Twitter. *Social Science Computer Review, 37(6)*, 721-739.

Mensier, A. (2023, May 29). *LLMs Could Become Weapons of Mass Disinformation* . Retrieved from Towards Data Science: https://towardsdatascience.com/llms-weapons-of-mass-disinformation-4def0dc3dc7

Mocanu, D., Rossi, L., Zhang, Q., Márton, K. & Quattrociocchi, W. (2015). Collective attention in the age of (mis)information. *Computers in Human Behavior, 51*, pp. 1198-1204.

Moreno, Y. and Nekovee, M. & Pacheco, A. F. (2004, June). Dynamics of Rumor Spreading in Complex Networks. *Phys.Rev. E, 69.*

Moscadelli, A., Albora, G., Biamonte, M.A., Giorgetti, D., Innocenzio, M., Paoli, S., Lorini , C., Bonanni, P. & Bonaccorsi, G. (2020, August 12). Fake News and Covid-19 in Italy: Results of a Quantitative Observational Study. *International Journal of Environmental Research and Public Health, 17(16).*

Myojung, C. & Nuri, K. (2020). When I Learn the News is False: How Fact-Checking Information Stems the Spread of Fake News Via Third-Person Perception. . *Human Communication Research. 47.*

Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve.* Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Okunoye, O.B. & Ibor, A.E. (2022). Hybrid fake news detection technique with genetic search and deep learning . *Computers and Electrical Engineering 103.*

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y. & Wang, W.Y. (2023). On the Risk of Misinformation Pollution with Large Language Models. *arXiv preprint arXiv:2305.13661.*

Pennycook, G., McPhetres, J., Zhang, Y. & Rand, D. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. *PsyArXiv [Working Paper]*, pp. 1–24.

Prieto, A., Prieto, B., Ortigosa, E.M., Ros, E., Pelayo, F., Ortega, J. & Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing 214*, 242-268.

Puraivan, E., Venegas, R. & Riquelme, F. (2023). An empiric validation of linguistic features in machine learning models for fake news detection. *Data & Knowledge Engineering 147*.

Quattrociocchi, W., Caldarelli, G., & Scala, A. (2014). Opinion dynamics on interacting networks: Media competition and social influence. . *Scientific reports. 4*.

Rahman, A. (2020, April 20). *Introduction to Genetic Algorithm and Python Implementation For Function Optimization.* Retrieved from Towards Data Science: https://towardsdatascience.com/introduction-to-genetic-algorithm-and-python-implementation-for-function-optimization-fd36bad58277

Rai, N., Kumar, D., Kaushik, N., Raj, C. & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering 3*, 98-105.

Ray, P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121-154.

Rosenblatt, F. (1958). THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN. *Psychological Review Vol. 65, No. 6*, 386-408.

Rubin V.L, Che Y. & Conroy N.K. (2016). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, Volume 52, Issue 1, 1-4.

Rubin, V., Conroy, N., Cornwell, S., & Chen, Y. (2017). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *NAACL-HLT*.

Ruffo G., Semeraro, A., Giachanou, A. & Rosso, P. (2022). Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review 47*.

Russell, S. & Norvig, P. (1962-. (2010)). *Artificial Intelligence A Modern Approach* . Upper Saddle River, N.J. :Prentice Hall.

Scheibenzuber, C., Neagu, L.M., Ruseti, B., Artmann, B., Bartsch, C., Kubik, M., Dascalu, M., Trausan-Matu, S. & Nistor, N. (2023). Dialog in the echo chamber: Fake news framing predicts emotion, argumentation and dialogic social knowledge building in subsequent online discussions. *Computers in Human Behavior 140*.

Schelling, T. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology, 1:2*, 143-186.

Shah, P. (2020, June 27). *Sentiment Analysis using TextBlob* . Retrieved from Towards Data Science: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

Shan, G., Wang, H. & Liang, W. (2019). Robust Encoder-Decoder Learning Framework towards Offline Handwritten Mathematical Expression Recognition Based on Multi-Scale Deep Neural Network. *arXiv preprint arXiv:1902.05376*.

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K., Flammini, A. & Menczer, F. (2018). The spread of low-credibility content by social bots. . *Nat Commun 9, 4787*.

Shu, K., Lee, D., Cui, L., Liu, H. & Wang, S. (2019). dEFEND: Explainable Fake News Detection. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395-405.

Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P. & Satoh, S. (2019). SpotFake: A Multi-modal Framework for Fake News Detection. *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*.

Sohail, S.S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., Atalla, S. & Mansoor, W. (2023). Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University – Computer and Information Sciences*.

Sun, R., Li, C., Millet, B., Ali, K.I.,& Petit, J. (2022, February). Sharing news with online friends: A study of network homophily, network size, and news type . *Telematics and Informatics, 67*.

Sunstein, C. (2002). The law of group polarization. *Journal of Political Philosophy, 10(2)* , pp. 175–195.

Tai, K.Y., Dhaliwal, J. & Shariff, S.M. (2020). Online social networks and writing styles–a review of the multidisciplinary literature. *IEEE Access 8*, 67024–67046.

Talwar, S., Dhirb, A., Kaurc, P., Zafare, N. & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services, 51*, 72-82.

Tambuscio, M., Oliveira, D.F.M., Ciampaglia, G.L. & Ruffo, G. (2018). Network Segregation in a Model of Misinformation and Fact-checking. *Journal of Computational Social Science, 1*, 261-275.

Tambuscio, M., Ruffo, G., Flammini, A., & Menczer, F. (2015). Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks.

Tchakounté, F., Calvin, K.A., Ari, A.A.A. & Mbogne, D.J.F. (2020). A smart contract logic to reduce hoax propagation across social media. *Journal of King Saud University - Computer and Information Sciences. 34.* .

Thiriet, C. (2023, April 20). *Large Language Models: Scaling Laws and Emergent Properties*. Retrieved from cthiriet: https://cthiriet.com/articles/scaling-laws

Thota, A., Tilak, P., Ahluwalia, S. & Lohia, N. (2018). Fake News Detection: A Deep Learning Approach . *SMU Data Science Review, 1 (3)*.

Financial Times. (2023, January 6). OpenAI: fundraising points to generative AI as next speculative bubble. *Financial Times*.

Trevisan de Souza, V.L., Marques, B.A.D., Batagelo, H.C. & Gois, J.P. (2023). A review on Generative Adversarial Networks for image generation. *Computers & Graphics, 114*, 0-12.

van der Linden, S., Roozenbeek, J. & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology, 11* , p. 2928.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomes, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *Advances in neural information processing systems, 30*.

Villa, G., Pasi, G. & Viviani, M. (2021). Echo chamber detection and analysis. *Social Network Analysis and Mining* .

Vosoughi, S., Deb Roy, D. & Sinan, A.S. (2018). The spread of true and false news online. *Science, 359*, 1146-1151.

Wang, Y., McKee, M., Torbica , A. & Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine, 240, article 112552*.

Wardle, C. (2017). Fake news. It's complicated. . *First Draft News*.

Wearden, G. (2023, May 26). Tech stocks surge as wave of interest in AI drives $4tn rally. *The Guardian*.

Wiesenberg, R. & Tench, R. (2020). Deep strategic mediatization: Organizational leaders' knowledge and usage of social bots in an era of disinformation. *International Journal of Information Management 51*.

Wolfe, C. (2022, December 10). *Language Model Scaling Laws and GPT-3.* Retrieved from Towards Data Science: https://towardsdatascience.com/language-model-scaling-laws-and-gpt-3-5cdc034e67bb

Wydmanski, W. (2022, December 3). *What's the Difference Between Self-Attention and Attention in Transformer Architecture?* Retrieved from Medium: https://medium.com/mlearning-ai/whats-the-difference-between-self-attention-and-attention-in-transformer-architecture-3780404382f3

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J. & Wen, J. (2023). A Survey of Large Language Models. *arXiv preprint arXiv:2303*.

Zhihan, L. (2023). Generative Artificial Intelligence in the Metaverse Era. *Cognitive Robotics*.

Zhou, J., Zhang, Y., Luo, Q., Parker, A.G. & De Choudhury, M. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-20.

Zhou, X., Pan, Z., Hu, G., Tang, S. & Zhao, C. (2018). Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets. *Mathematical Problems in Engineering*.

Zi, C., Gianvecchio, S., Wang, H. & Jojodia, S. (2010). Who is tweeting on Twitter: human, bot, or cyborg? *ACSAC '10: Proceedings of the 26th Annual Computer Security Applications Conference*, 21-30.