LUISS

Corso di laurea in Data Science and Management

Cattedra Machine Learning

Evaluating the impact of tweet characteristics on audience engagement with fact-checking: a machine learning approach

Prof. Giuseppe Italiano

RELATORE

Prof.ssa Irene Finocchi

CORRELATORE

Daniele Fiorucci (Matr. 761461)

CANDIDATO

Anno Accademico 2023/2024

Ai miei genitori,

coloro che mi hanno insegnato il valore del sacrificio

senza mai avermi fatto mancare il loro supporto.

Abstract

This study aims to analyze the impact of the characteristics of fact-checker tweets on social network users, focusing in particular on those who do not follow these pages because they are considered more vulnerable to misinformation.

The thesis is developed around large datasets of tweets focused on six fact-checking pages from the UK, Italy and France, obtained via Twitter's API and extended using data engineering practices.

In a historical context in which technologies, innovation and especially artificial intelligence (AI) are making great strides, the risks associated with the dissemination of artificial content are there for all to see.

By leveraging machine learning models, this research addresses a critical gap in the literature, offering data-driven recommendations focused on fact-checkers, to improve their presence on social networks, a key medium for news dissemination nowadays.

In particular, it seeks to exploit elements such as sentiment, the number of hashtags, emoji, links, mentions and the number of words in the tweets published by these pages in order to reach new users.

The ultimate aim is to make the fight against disinformation even more effective and to contribute to more informed public information.

TABLE OF CONTENTS

1. Introduction

Social networks are online platforms that allow users to connect and exchange different types of content with each other.

Although each social network has specific characteristics, the most popular of which are Twitter (now known as X), Facebook and Instagram, they all share common features:

- Profile customisation: a user's personal page containing photos, personal information, relevant content, experiences, interests
- Follows and connections: means of connection to other users, may be unilateral or bilateral and may require prior acceptance by the other party before acceptance
- Sharing of material: sharing, depending on the privacy chosen, of blocks of text, photos and videos, only to certain users or freely to the whole network
- Creation of communities: aggregation of users with common interests or objectives. The formation of communities can occur either automatically or through user initiative. In some social networks, communities are explicitly defined by the platform and can be moderated by one or more users who establish the rules for content sharing. In other contexts, the term “community” refers to groups of users who interact with each other without an official structure.

The element that plays, however, a fundamental role in social networks is the recommendation algorithm on which it is based.

The algorithm aims to show each user, in addition to the content of the people in their network, unexplored content that they find affinity with based on their past actions. This is crucial to provide users with personalised experiences.

Another fundamental phenomenon of social networks, closely related to the recommendation algorithm, is that of viral content.

The viral content system is based on several elements that lead to rapid spread within the entire social network. This occurs when a large number of users re-share a certain material, an effect that is accentuated if the re-sharing is carried out by particularly influential users. In addition, once a piece of content has gone viral, this will most likely also be suggested by the recommendation algorithm, so viral content can reach an unprecedented echo.

Moreover, often a matter of concern from a protection and privacy point of view, social networks record users' personal data, so as not only to target content, as mentioned above, but with the aim of providing specific advertisements, this being the main source of revenue from social networks.

However, this digital revolution presents both lights and shadows.

First of all, the possibility of publishing news, either anonymously or using a name that does not correspond to the real one, makes it impossible to verify the credibility of a piece of content based on the reliability and background of the person sharing it, nullifying the possibility of carrying out an initial check, a common practice in traditional means of disseminating information. One of the main negative aspects that emerges from the dissemination of news on social networks is the lack of a verification process, which in traditional media helps to ensure the accuracy of information before it is published.

Another problem is related to one of the main characteristics that led to the diffusion of social networks, and which also applies to the content shared on them: the positive network effect. Positive network effects occur when the value of a service or product increases as more people use it (Kornish, 2004). In the context of social networks, this means that the more people start to share and reshare information, the more users will see and engage with it.

This characteristic of social networks amplifies the viral nature of the content, including tweets. Viral content spreads rapidly and widely, often beyond the original audience, because each reshare or retweet introduces the information to a new set of users. As a result, tweets that start to gain traction can quickly reach a massive audience, even if the contents are only partially true or not true at all. This amplification can lead to misinformation spreading quickly, making it difficult to discern the accuracy of viral news.

Furthermore, the introduction of deep fake for image and video synthesis (Demuyakor, & Opata, 2022), as well as generative artificial intelligence, has meant that nowadays, unless the information is studied before determining its truthfulness or falsity, it has become so credible that it can mislead anyone.

The fact that in order to defend oneself against false information, it is essential to verify sources and information, a time-consuming and complex process, is necessarily at odds with the assumption that has led to the success of social media as a source of information, the time-saving.

In order to combat the spread of fake news on social media, groups and organisations have sprung up with the aim of confirming or disproving the news that has picked the public's interest, their name is fact-checker.

The methodologies through which fact-checkers operate are diverse and depend on the intrinsic characteristics of the organisation. They can involve the use of innovative software or tools, as well as be carried out entirely by human beings.

Among the most innovative solutions is certainly the one concerning the use of artificial intelligence for the self-detection of fake news.

The use of AI, through advanced models of machine learning, is certainly a revolutionary solution, especially since it would reduce the long time it takes humans to carry out accurate fact-checking, but it still has many limitations nowadays. This solution has some fundamental problems concerning the difficulty for the machines to perceive sarcasm and the delay in learning new facts, since the model is trained only once and is not updated in real time. This implies that it cannot quickly adapt to new information or changes in the context of the news. Furthermore, as Hao (2018) observes: “Even the best AI for spotting fake news is still terrible.".

Moreover, news cannot simply be labelled “true” or “false”, but there are so many facets it can take on (Molina et al., 2021). This blurring of boundaries makes the classification task of machine learning or AI algorithms even less easy.

Instead, manual fact-checking is usually performed by journalists who manually carry out the control and study of news, especially viral ones, with the aim of disproving or confirming the news. These latter, to combat the positive network effect that leads fake news to become viral, need their own profiles and their own publications to become viral themselves.

Manual fact-checking, although costly in terms of time, is certainly the safest method of verifying information there can be. We can therefore ascertain the fact that a user who is familiar with the fact-checking page and constantly visits it, translating this into social media terms we can say that he or she “follows” the page, is protected by all the information that passes through the page.

On the other hand, there are still many users who, even though they use social networks on a daily basis, do not verify information and do not even follow fact-checking pages.

These are the ones who are the most vulnerable and susceptible to fake news, the users who, due to personal convictions or uncritical acceptance of information, do not or are not able to carry out an independent check of the news.

These are the users that fact-checkers aim to reach, and the users for whom fact-checkers can be a truly crucial tool.

How can, therefore, fact-checkers reach the profiles of users who still do not follow these pages? What should be the characteristics of tweets to expand their circle of users?

This is the main question that this work aims to answer, through statistical and machine learning techniques, as well as decidedly innovative tools such as gen AI and keyword extraction.

This is because technology, and increasingly powerful computational capabilities, will be fundamental tools to counteract the spread of fake news, but as mentioned they are only a tool in the hands of real people, since a tool that can independently identify fake news is not yet possible.

1.1. Why is this problem important?

Fake news finds fertile ground in this period of great technological innovations that make them more and more credible, increasing the difficulty for users to independently refute or confirm the news they learn on social networks.

The consequences of this are first and foremost the spread of distrust in the media, including traditional media, and therefore in institutions. This mentality, which has been widespread for some years now, leads to social non-cohesion and the emergence of conspiracy theories, which are not the root of fake news, but the direct consequence.

Furthermore, the study of Ecker et al (2022), shows that even after denying or confirming false news, psychologically speaking, the person continues to be influenced by it. This further emphasises the importance of minimising the time between the publication of fake news and its correction.

This manipulation of popular opinion can influence a democratic process such as elections, through the dissemination of misleading news aimed at manipulating voter opinion, as has happened in the past.

In relation to the 2016 US elections, in which the term “fake news” was pronounced for the first time, it has been shown (Allcott, & Gentzkow, 2017) that many US adults believed fake news shared via Facebook, and that most of them would favour Trump.

Only two years later, a similar event occurred in the 2018 Taiwan election (Wang, 2020). In this case, it was shown that neutral voters tended to vote for the candidate wanted by the Chinese government more easily, driven by fake news and the inability to distinguish it from real information.

These three real analyses show that fake news and misinformation represent a real danger to democracy, as they are able to influence the public opinion of people who do not have the methods or are not able to verify the reliability of news, something that has become increasingly complex due to technological innovations capable of generating highly credible fake news.

1.2. Literature review and identification of research gaps

Fact-checkers play a key role in the battle against disinformation, on social networks and beyond. The methods used to examine the news, as well as the topics they deal with, depend on the intrinsic organisation of the fact-checkers themselves, they can be profit or non-profit as well as being born as supplementary services of newspapers, such as CheckNews, which originates from the French newspaper Liberation, they can be entities not connected to any type of newspaper as well as being born as blogs and then becoming actual fact-checkers.

Although fake news has always existed even before major technological innovations, it saw substantial growth during the 2016 US election (Allcott, & Gentzkow, 2017), in which they found the new social media platforms to be the perfect conductor.

In parallel with the rapid increase of fake news, fact-checkers have also grown in importance and have achieved international reach and relevance over the past years, although of course depending on the country and social context in which they arise, they have faced different challenges of varying magnitude, as described in the work by Laurens and Graves (2021).

As mentioned above, social networks have facilitated an unprecedented free accessibility of news. On the other hand, the checking of news before it is published cannot be compared to that of paid newspapers, which offer a quality service in this respect.

Fact-checkers plays now a leading role, not only in denying or confirming the news, but also from the point of view of studying the algorithm of social networks, so as to understand which news is, or will become, viral, in order to decrease as much as possible the number of people who, unable to fact-check independently, fall trapped in fake news (López-Marcos & Vicente-Fernández, 2021).

The impact of fact-checkers in social networks is demonstrated by several studies including the one by Bachmann and Valenzuela (2023). This aims to determine whether fact-checkers improve or worsen the reputation of newspapers and media, so as to decrease distrust of them by citizens, one of the most important consequences of the spread of fake news on social networks. Furthermore, the study written by Ecker et al (2022), shows that even after fake news is disproved, it continues to influence people's reasoning.

Other studies, such as the one by Dafonte-Gómez et al (2022), explain the presence on social media and the use of content distribution tools by fact-checkers from different parts of the world. As we will see later, our study also aims to generate insights so that fact-checkers' posts can reach as many new users as possible, albeit with a totally different approach.

Obviously, the moment when popular opinion becomes even more important than everyday life is when it can decide the fate of a country, and not only that, during elections.

The creation of fake news, which began its golden age precisely with the 2016 state elections, has assumed a fundamental role, in a negative way, in several elections that have followed one another around the globe, not stopping with the American ones of the last decade (Allcott & Gentzjow, 2017; Wang, 2020).

In addition to the two sources already mentioned in the previous paragraph, the work by Allcott and Gentzkow (2017), which specifically delved into the role and impact of fake news in the elections that saw Donal Trump triumph in America, and the pro-Chinese candidate in Taiwan (Wang, 2020), there are also other cases where fake news has tried to affect elections.

Most notably, the study written by Cazzamatta and Santos (2023), shows the categories of news and their characteristics that attracted the attention of the main fact-checkers during the 2022 elections in a country where disinformation reigns supreme like Brazil.

In the field of fact-checking, there are several methodologies and approaches that can be used, each with its own pros and cons.

On the one hand, the use of technology significantly reduces the time required for processes that can take hours, if not days, when performed by a human being (Hassan et al., 2015). On the other hand, these are still inaccurate as they lack the sensitivity and critical thinking typical of human beings, which in fields such as Fact-checking can be crucial.

The literature, in this case, does not take a definite position but goes on to analyse the characteristics of each of these schools of thought.

- **Automated fact-checking models**

  An example of the use of Natural Language Processing (NLP) regarding fact-checking for the automatic identification of fake news is described in the work by Guo et al (2022). Here, an entirely AI-driven framework is proposed that consists of 3 stages, the first being “Claim Detection”, followed by “Evidence Retrieval” up to the production of the output that consists of a “Verdict Prediction” and a “Justification Production”.

  A comprehensive overview of the pros and cons of what is called AFC (Automated Fact-Checking) is provided in the work written by Graves (2018). While again identifying what could be the pipeline of the process, divided into "Identification", "Verification" and "Correction", the study lists what are the limitations that make AFC something still far from reality. In fact, through interviews with fact-checkers and computer scientists, it emerges that at the moment the idea of fully automated fact-checking is something far away, and that at the moment it can only act as a support to what must still be a process carried out by human beings.

- **Hybrid solutions**

The paper written by Nakov et al (2021), one of the co-authors of which is a Senior Data Scientist in one of the fact-checkers pioneering the use of software to support human work, namely FullFact, one of the fact-checkers used in our analysis, speaks precisely of a hybrid solution.

The cited paper explores what functions can be performed by AI to reduce the workload on humans, the main ones being:

- Identifying news stories that deserve to be verified, trying to anticipate those that have the potential to go viral
- Finding relevant claims that have already been verified in the past as true or not
- Finding supporting evidence
- Translating, transcribing and summarising textual and non-textual content

By implementing these AI-driven functions, the fact-checking process becomes more efficient, allowing human fact-checkers to focus on more complex verification tasks and ultimately improving the accuracy and speed of fact-checking.

- **Fact-checking made by humans**

Finally, the methodology that is certainly the most effective, but which possesses insurmountable problems in terms of scalability, is that of fact-checking carried out by human beings without support from technology.

There are several methodological pipelines that can be used, which vary among fact-checkers. We will now specifically mention one fact-checker taken into consideration in our study, which shares the characteristic that the verification is carried out entirely by human beings.

This is the Italian fact-checker that deals mainly, if not entirely, with the checking of political news, this is Pagella Politica. As described in detail on their website (https://pagellapolitica.it/progetto), a 10-step process is used, reviewed by multiple professionals before publication.

There are also studies that conceive the idea of human fact-checking with alternative methodologies, a clear example is the paper written by Allen et al (2021). This study examines the potential of crowdsourcing, made up of politically balanced, ordinary people, to identify misinformation, concluding that this approach can also provide accurate assessments comparable to those carried out by professional fact-checkers.

A broader analysis of the audience most likely to be reached by fact-checkers was analysed, with reference to the United States, in the paper written by Robertson et al (2020), which leads to the conclusion that beyond the specific algorithms of different social networks, the use of fact-checking sites is also influenced by the ideological perspectives of the users, and the users' news consumption tendencies.

Other analyses, such as the study written by Lim (2018), highlight the difficulties in achieving an objective and unambiguous assessment of different fact-checkers, when news cannot be assessed as completely true or false, adding further to the risk that strategic responses from politicians may fuel this subjectivity of interpretation.

There are, therefore, several studies that show the different methodologies that can be used to carry out fact-checking, i.e. automated through the use of technological innovations, entirely man-made and hybrid, as well as the audience that can more or less easily be reached by fact-checkers and the considerable importance of fact-checking linked to the social network environment.

Precisely within social networks, as already mentioned, in order to counter the fact that fake news can spread, it is crucial that the posts of fact-checkers also go viral, or reach as many users as possible. What is missing, therefore, is a guideline according to which fact-checking accounts should generate content on social networks, so as to inform as many users as possible.

This study aims to provide a data-driven analysis carried out on a large number of contents published on social networks, Twitter specifically, by six different fact-checkers. The result of the analysis will provide insights into the characteristics of the content that can most effectively reach those who do not yet follow these accounts.

Through this study, fact-checking accounts will be able to build their strategy to reach the widest possible audience in order to make the most of their work.

1.3. The present study

The aim of this study is to provide data-driven insights as the output of an ad-hoc study carried out on fact-checker profiles on social networks.

Starting from our work, fact-checking pages will be able to elaborate their communication strategy in such a way as to reach users who do not yet follow the fact-checking page in order to expand their user base.

Our study therefore aims to close the research gap illustrated above by answering the following research question:

- What features do tweets need to have, such as sentiment, number of hashtags, emojis, links, quotes, mentions, and word count, to effectively reach new users?

Our aim is not to identify fake news, as many studies in this field do, nor to identify promising news that could go viral, because it is placed at a later stage than the classification of content, in fact, it can be used either by humans or by technology. This work aims to help fact-checking pages at the moment the content is published.

By means of this study, which is a social network analysis carried out on data specifically concerning only tweets and fact-checking pages, we are going to study what characteristics a tweet should have to generate impressions on all users in general, and especially on those who do not yet follow the page.

Although this phase of the fact-checking processes, i.e. that of publishing content verification, is not covered in depth by the literature, without it all the other studies on models and algorithms for recognising fake news prove useless.

Indeed "The effectiveness of the work carried out by these journalistic initiatives depends not only on the quality of their content, but also on their ability to reach large audiences through the same channels by which disinformation spreads." (Dafonte-Gómez et al., 2022).

We therefore analysed 6 fact-checkers from 3 different countries, to provide a study that could be as universal as possible and generalisable to any other fact-checker.

For each fact-checker, the datasets contain not only information regarding the intrinsic characteristics of the tweets, such as the number of words or emojis contained, but also information regarding the number of retweets, both from page followers and not-followers, and additional considerations such as the sentiment analysis of the tweet, extracted through generative AI models, as we will go into later.

Starting from this data, after a data engineering phase to extract new hidden information, we carried out the data exploration. Here we studied the general characteristics of the tweets, such as the average number of words used, the number of emojis, links and hashtags, the sentiment of the most common tweets and the type of tweets whether retweets, repeats or original tweets. Above all, a comparison of each of these characteristics was implemented for each of the different tweets and also a comparison between the tweets themselves, this phase is important because it could generate essential considerations when studying the final results.

We then carried out the independence tests between the features that we decided to take into consideration for the purposes of the study. We carried out chi-square tests between the qualitative variables, the correlation matrix between the quantitative ones and finally the Kruskal-Wallis test between one quantitative variable (“impression”) and some of the qualitative variables.

Finally, we arrive at the most important phase of the study, the modelling section. A machine learning model was selected and then used to specifically analyse information concerning non-followers who nonetheless shared tweets from the fact-checking page.

To obtain this type of analysis, we had to open the black box of our machine-learning model, in order to make the results readable and interpretable.

The model chosen, the log-linear Poisson regression, is not as complex as many others, but achieving complexity has never been our objective. Rather, our aim is not focused on the predictive capacity of the algorithm, but on the possibility of obtaining clear results that can be transformed into comprehensive reports. These reports provide detailed insights into the impact of each feature of the tweets on the potential to reach users who do not yet follow the fact-checking page.

2. The new frontier of fake news: the deepfake

Technological innovations, as discussed above, supported by the development of revolutionary machine learning algorithms, rapid increase in the amount of data generated and increasingly powerful computers, may in the not-too-distant future represent the most effective weapon against fake news and disinformation.

On the other hand, this technological revolution reserves lights and shadows.

These technologies can, in fact, be used as a tool to spread fake news and misinformation, with the aim of orienting public opinion, or simply discrediting the figure of a person, through the use of innovative techniques that make detection more and more complex, and this is how the phenomenon of deepfake was born in recent years.

Deepfakes are innovative technologies that allow the creation of highly realistic images, videos and audio.

Deepfakes are based on algorithms such as deep neural networks (deep learning) and the Generative Adversarial Network (GAN), artificial intelligence (AI) techniques that allow the manipulation and generation of multimedia content.

It is therefore possible to generate videos in which politicians or celebrities appear to do and say something that never really happened, spreading disinformation, and influencing public opinion. From an audio perspective, voice synthesis technologies can be used to replicate a person's voice or simulate conversations.

In sensitive contexts such as wars or political elections, deepfakes can have serious consequences that can really test the world balance and democracy, as well as citizens' trust in institutions and technology.

It is necessary to emphasise, as done in the study by Vaccari and Chadwick (2020), the power of information disseminated through audiovisual material.

Especially in the political sphere, this type of content has more impact than written content, as demonstrated in the study by Graber (1990). Further proof comes from the study by Stenberg (2006), who showed that humans have an easier time metabolising and remembering video content than written content, this phenomenon is called the “picture superiority effect”.

Deepfakes thus exploit the inherently persuasive nature, due to the stimulation of multiple sensory apparatuses of the human being, of video content and sound to disseminate content.

The spread of deepfakes leads to the emergence of new challenges, both from a technological as well as an ethical and legal point of view, for the protection of people (Khan et al., 2022).

Deepfakes pose a real risk to democracy, as well as to the reputation of individuals, as these techniques can be used to influence elections and the reputation of political figures, creating false scandals as we shall see later.

Fundamental, is the use of these same technologies to recognise and attempt to counter the disproportionate growth of these phenomena, so that artificial intelligence can become a tool to protect information, transparency and truthfulness, and not an obstacle to be feared.

In addition to the technological race between those who use technological innovations to produce deepfakes and those who instead use them for the detection and countering of them, it is also crucial to promote, through education, a system of public awareness that can provide citizens with the fundamental knowledge to critically evaluate media content.

Technology specific knowledge, and thus teaching the basics of the technologies that can be used to create deepfakes, should help open the black box behind the magic that enables the creation of deepfakes, developing a less pessimistic and more critical attitude (Meskys et al., 2020).

Digital and media literacy must enable citizens to be educated about the risks of using technology to distort information, as well as educate them about the existence and use of tools such as fact-checking platforms, which will be an increasingly crucial resource.

2.1. Fake news, deepfake and social network

Deepfakes can be seen as the evolution of fake news, which has recently reached levels of persuasion that were unimaginable until recently.

Whereas traditional fake news was typically spread through simple texts or at most images, which can often be detected without too much effort, today deepfakes represent a new and innovative method of spreading fake news. These use cutting-edge technologies to convincingly replicate reality by distorting audio or even video.

This new technology sees social networks as a medium that fuels its reach, in fact, those who use these technologies use social media to amplify the impact of content, and misinformation, in society.

This is because the platforms have, as seen above, the pillar of maximising the engagement of as many users as possible.

In addition to the inherent characteristics and purpose of social networks, user behaviour also plays a key role. The study conducted by Pancer and Poole (2016) showed that during the 2016 presidential campaign, Trump and Clinton's tweets that contained images or videos received significantly more re-shares and likes.

Deepfakes are, therefore, not only a problem in terms of detection and greater credibility than the written text, but they also match the behavioural preferences of users, increasing, even more so, their spread.

However, returning to the discussion of the lights and shadows of technological innovations made earlier, while the sharing algorithm and architecture of social networks enriches the user experience, it also facilitates the spread of fake news.

Moreover, social networks do not possess effective mechanisms for verifying the veracity of information, as is the case with newspapers, which is why fact-checking plays a key role and this study represents a tool they can use to maximise the reach of their content.

A discussion regarding the role of social networking platforms against deepfake-generated content is increasingly topical, as highlighted in the study conducted by Meskys et al (2020).

One response was in 2019 from Facebook, the company in collaboration with Microsoft launched a competition, with a monetary prize of several million to the winner, whose aim was to build the best possible model for deepfake detection (Dolhansky et al., 2019).

The role and tasks to be played by social networks in combating the disinformation caused by deepfake, however, is much more complex and still generates wide-ranging debates.

On the one hand, there is the obligation of platforms to implement technologies aimed at protecting users, attempting to filter and remove deepfakes that might influence public opinion.

On the other hand, an overly strict intervention could have the opposite effect, resulting in a form of censorship that could limit users' freedom of expression (Khan et al., 2020).

As in many situations, the optimal approach is a balanced solution that is thought out in detail through collaborative supervision between humans and technologies and legal transparency from a policy perspective.

The combination of new techniques for creating fake content, deepfakes, and the sounding board of social networks can have particularly serious consequences when content is utilised during critical events, to which the public is particularly sensitive, such as elections or wars.

Indeed, among the innumerable risks associated with the deepfake phenomenon, among the main ones we can identify the damage caused to the reputation of individuals, the manipulation of public opinion and the breakdown of trust in institutions and authorities.

2.2. Exploring deepfake risks: politics, wars and pornography

Deepfake technologies can be applied in a variety of ways ranging from face swapping, and video generation to lip-syncing fake audio.

A practical example of use in politics that occurred in 2018, created by Peele and BuzzFeed, created to raise awareness of the evolution and uses of technology, shows former President Barack Obama uttering sentences that were never said, describing Donald Trump as a “total and complete dipshit” (BuzzFeed/YouTube, 2018).

The example of the fake video about the former president dates back a few years, but in the most recent period there are more and more, and better quality, uses of these technologies to influence public opinion and manipulate information.

The study conducted by Dobber et al (2021), provides, from a psychological point of view, an analysis of an experiment carried out to examine how deepfakes in the political sphere influence attitudes towards certain personalities and/or parties.

It is shown that the use of deepfakes towards specific groups (microtargeting) is able to enhance effects on political perceptions.

More recent examples concern the use of deepfakes in contexts of war, such as that between Russia and Ukraine.

As extensively analysed in the study conducted by Twomey et al (2023), there are several cases in which deepfakes were used for both sides with the aim of influencing international public opinion, the mood of soldiers on the war field and the geopolitical situation.

The most prominent examples involved fake video messages from high-ranking officials, such as a message showing Russian President Vladimir Putin at the moment of announcing a fake peace, or another portraying Ukrainian President Volodymyr Zelensky in the act of declaring his country's surrender.

While a physical and positional war is waged on the ground, a parallel conflict unfolds in the digital realm. This technological warfare is primarily informational and strategic, focused on the capability to generate confusion and disinformation, thereby altering the perception of real events.

Another type of deep learning application is that of a purely defamatory nature, which in the majority of cases concerns the production of fake explicit porn videos, whose actors' faces are replaced with those of other people without their consent.

According to a study by Deeptrace: “96% of deepfake content was found to be pornographic in nature” (https://www.bbc.com/news/technology-49961089). Furthermore, a report by The Guardian clarifies that, of these pornographic contents: “99% of those mapped faces from female celebrities on to porn stars.” (https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them).

The objectives of these videos can be multiple.

On the purely defamatory side, as in the case where the face applied in the videos is that of a celebrity, a high-quality video that is difficult to distinguish from the real thing may represent an image damage that takes time to rectify.

Another objective may be revenge, which would then configure the generation of explicit material through deep learning techniques as a new form of porn revenge.

One of the weaknesses of institutions that have emerged as a result of the rapid technological development is the inability to keep up from a regulatory point of view, unable to offer an adequate and common response (Mania, 2024).

Another time when deepfakes represent not only a means of media manipulation, but also a danger to democracy is elections.

As explored in the work of Nieweglowska et al (2024), deepfakes are particularly effective, not only as a means of persuasion, but also as a means to radicalise individuals, as these technologies leverage on already existing extreme beliefs, but which are pushed to extreme levels.

2.3. Types of deepfakes

- **Generation of original faces:** the creation of completely new faces through the use of Generative Adversarial Networks (GANs). The image is generated via a generative network, after which its authenticity is judged via a discriminator. The result is a multimedia content, which is difficult to judge as synthetic because human characteristics are reproduced in a very realistic manner (Bansal & Joshi, 2021).
- **Face Swap**: a methodology consisting of the exchange of faces between two or more subjects within an image or video. This technique has been practised in the past by means of graphical editing software methods, recently, machine learning algorithms have been developed that can perform the task automatically and with greater quality. Amongst these, one of the most advanced is FSGANs (Face Swapping GANs), a GAN-based technology that enables face swapping without the need to train the model on specific individuals, reducing costs from both a data collection and timing perspective (Nirkin et al., 2019).
- **Face Reenactment**: this technique involves the transfer of facial expressions and movements from one video to another (Bansal, 2021). An innovative model to achieve this is the AVR-GAN (Audio-Visual Face Reenactment), a GAN-based technique, proposed in the paper by Agarwal et al (2023). This involves not only the use of different

videos as input but also the integration of audio. This innovation makes it possible to synchronise lips and movements, combining audio and video sources to make animations more natural.

- **Manipulation of facial attributes**: SC-FEGAN (Sketching Composite - Face Editing Generative Adversarial Network), a methodology based on GANs, was introduced in the study conducted by Jo and Park (2019). This allows the details of faces in images to be modified through graphical inputs, such as design and colours, set by the user, allowing images to be customised according to the user's wishes.
- **Voice cloning**: a technique that allows any text to be read with the replicated voice of a given person (Khan et al., 2022). This is one of the most frightening techniques from a security point of view, as the result is among the most difficult to detect products and content generated with deepfake technologies.
- **Multi-Modal Techniques**: techniques that combine multiple types of multimodal content, such as audio and video, to create more realistic deepfakes than ever before. The use of these technologies allows for very high levels of realism, and the fact that multiple senses of the user are captured in a believable way increases user immersion (Khan et al., 2022). Generative techniques based on Deep Learning, have enabled the creation of increasingly sophisticated content and, above all, more and more easily (Liz-Lopez et al., 2024).

2.4. Deepfake generative's models

- **GAN (Generative Adversarial Networks)**: the GANs are a type of machine learning algorithm consisting of two main elements, a generator and a discriminator. The former has the task of creating the media content, while the discriminator has to assess whether these are false or can be traced back to reality. This is an iterative process that allows a constant improvement in the quality of the generated content as the generator-discriminator cycles proceed. These technologies present a technical challenge given the unstable training that can lead to "mode collapse", in which the generator produces a limited amount of different outputs, putting a brake on the realism of the results (Chauhan et al., 2024).
- **Autoencoder**: among the first deepfake generation models, they learn to encode (encoder) the input into a compressed object and then decode it (decoder) to construct

the output. The goal of training these types of models is to minimise the error difference between the input and the output (Patel et al., 2023). This methodology consisting of encoders and decoders is crucial for generating very high-level deepfakes that are difficult to distinguish from the originals. Autoencoders and GANs both aim to generate deepfakes, although the former are more suitable for unsupervised learning and dimensionality reduction, while the latter emulates an existing distribution (Khan et al., 2022).

- **Variational Autoencoders (VAE)**: types of generative models that belong to the class of autoencoders. VAEs, possess, like classical autoencoders, an encoder and a decoder, but unlike the others, these possess a probabilistic approach (Jo and Park, 2019). The encoder learns an encoding that maps the input to a distribution of possibilities and then uses this distribution to generate an output as close as possible to the input. VAEs are a powerful tool in generating images from some basic realism (Zendran & Rusiecki, 2021).

  There is also an alternative approach, VAE-GAN, which aims to produce content that is increasingly close to reality. This combines VAE with a generator instead of a decoder and a discriminator that classifies real images from generated ones (Patel et al., 2023).

- **CNN (Convolutional Neural Networks)**: CNNs are a type of deep learning algorithm that is particularly applicable for image analysis and image processing. These particular neural networks are designed with the aim of learning the different patterns of images autonomously, a fundamental feature for generating content that can replicate reality in the best possible way. The main limitation of these algorithms is, however, the need for a large amount of data in order for the training of the model to be effective, as well as a large amount of computational power. Once trained, the latter can be used to generate as many deepfakes as desired. For the synthesis of this content, another generation model is used, often a neural one such as GANs.

  In addition to the point of view of deepfake generation, CNNs are also a very valuable tool for the detection of fictitious media content.

- **Autoregressive models**: autoregressive models are machine learning models that are based on the sequential generation of data, they predict the next output based on previous inputs and continue iteratively. In the context of deepfake synthesis, autoregressive models predict each frame based on previous frames. The sequential approach is particularly effective in detecting movements and expressions of figures, as transitions appear smooth and ensure high-level temporal consistency (Khan et al., 2023). As with CNNs, training this type of model requires a large amount of computational resources.

2.5. Fact-checking and deepfakes

The generation of completely fictitious images, videos and audio, i.e. the deepfake, poses significant new challenges to those who monitor content on social networks to counter misinformation, the fact-checkers.

The improvement in the quality of deepfakes, due to technological innovations, has certainly added more than one layer of complexity to those who have to sift through content. Fundamental to combating new technologies is knowing them, a concept also expressed in Engler's study (2019). There are several reasons why it is of considerable importance for journalists to know the algorithms and models involved in the generation of media content.

First of all, so that they can learn what the limits of these innovations are in order to understand what the alarm bells may be that signal deepfakes. Secondly, understanding the technology allows them to be ready when new forms of deepfake may appear. Thirdly, fact-checkers can collaborate with lawmakers to mitigate the risks associated with deepfakes in an active way. Finally, they can play an active role in educating and preparing the public, not only by declaring when a piece of content is fake, but also how to recognise it, since they cannot handle all fictitious content and the public can begin to do so themselves.

Other models and tools that are seeing growth, both in number and effectiveness, in parallel to the deepfake generation tools, are those concerning the detection of altered or synthetic content.

Given the complexity and quality of the new deepfake technologies mentioned above, it is important that in parallel to a study of journalists regarding the credibility of information on the basis of further topical and political contexts, there is also, and not entirely, use of automatic software for the detection of artificial intelligence generated content.

Another fundamental element is that of collaboration between the various bodies and organisations involved in fact-checking.

Indeed, collaboration allows the possibility of sharing software and resources, increasing the ability to identify deepfakes, especially between countries.

Moreover, by collaborating internationally, it is possible to share information about different historical and cultural contexts, speeding up the process of verification and effectiveness.

In conclusion, cooperation would facilitate the possibility of sharing detection tools and information, allocation of resources, acceleration of processes, and would also foster greater geographical coverage in the battle against deepfakes.

2.6. Regulation and education

Therefore, on the one hand, it is important to continue to exploit the evolution of technological innovations as a means of combating disinformation, on the other hand, it is crucial to see and know what the evolution and challenges posed by deepfake may be.

Precisely in this regard, as with all innovative technologies, the need to standardise laws for the protection of affected citizens has grown.

The study by Meskys et al (2020) highlights precisely the parallel need for regulations and, from an ethical point of view, to keep up with this technological revolution that has been seen in recent years.

From the point of view of legal protection, there is a need for the emergence and uniformity in Europe as well as worldwide of a protection framework that would define a new type of crime or associate it with an already existing type.

In fact, in the study conducted by Mania (2024), it is described how many national legislations are still not equipped to counter technological innovations that can lead to the abuse of online media content. In particular, the need is emphasised to combat porn deepfake, protecting individual victims, and even more so applications of deepfake in the political sphere or concerning sensitive topics (e.g. wars), protecting democracy as a whole.

To be able to counter the inappropriate use of new technologies for purposes that can harm the community, it is also essential to educate others in the correct use of innovations. This is because, as described in the paper by Widder et al (2022), the problem of ethics emerges with regard to open-source technologies. It is described how open source communities take for granted the negative implications necessary for progress, a theory and mindset that could, however, lead to underestimating the consequences of new innovations.

It is crucial to work on combating deepfake from a prevention perspective that is not only legal but also ethical, fostering not only the principle of transparency at the heart of open source communities, but also moving in an ethical direction so as to anticipate the harms of these technologies.

2.7. Collaboration between different actors

The optimal solution for countering and combating deepfakes, and more generally disinformation caused by technological innovations, must consist of a holistic approach combining technology, ethics and legality.

First of all, laws must evolve in step with the evolution of technologies, protecting against defamation and the right to privacy. A quality regulatory framework must also aim to encourage the detection of deepfakes by promoting the responsibility not only of users and creators of fake content, but also that of social networks.

Secondly, algorithms and models must be exploited, in collaboration with human control (fact-checking), for the rapid and effective identification of altered or synthetic content. Once deepfakes are identified, the disclosure of the outcome by fact-checking organisations, especially on social networks, is crucial.

This is the context of this study, which aims to provide the latter with an ad-hoc tool that can be used to reach not only the largest number of accounts, but those that in particular do not yet follow fact-checking pages. In order to achieve this, the analysis we offer clarifies which tweet characteristics can be acted upon.

Last, but by no means least, is the need for educational programmes accessible to all with the aim of promoting digital literacy and continuous updating in order to stay abreast of technological innovations.

The creation of this ecosystem through the active collaboration of all the actors mentioned above is the only possible way to ensure a more transparent, secure and reliable information environment.

3. Methods and Analysis

The main purpose of this chapter is to provide an exhaustive description of the datasets used for the research, the statistical and machine learning techniques employed, the pipeline on the various steps performed during the preparation of the data, and the analysis carried out.

Initially, an in-depth description of the datasets, their collection and characteristics will be provided.

Next, the techniques used will be described in detail, and the justification of why they were chosen over others, in relation to the characteristics of the data at hand and the final objectives of the analysis.

After laying the foundations from a conceptual and methodological point of view, the actual research will be examined in detail, carried out concretely using the Python programming language. Using this tool, it was possible to manipulate, process and visualise data in a transparent and, above all, efficient manner.

3.1. Dataset and schema

During the analysis, we studied the Twitter profiles of six different fact-checkers, with the aim of understanding which tweet characteristics positively influence users' engagement, particularly those who do not follow the pages of the fact-checkers we considered.

The fact-checkers selected, according to principles that we will discuss later, come from different countries. FerretScot and FullFact from the United Kingdom, PagellaPolitica and FactaNews from Italy, AfpFactuel and CheckNewsfr from France.

The choice to study six fact-checkers from three different countries is twofold.

Firstly, the choice to consider fact-checkers operating in different countries made it possible to cover a broader geographical and cultural spectrum, offering a global vision that would allow us to go beyond the borders of a single country, but making the results of the analysis consistent to generalise the conclusions reached as much as possible.

Secondly, we chose not to consider only one fact-checker per country, but two, reducing the probability of an outlier within the country, especially if the results of the analysis are repeated in both fact-checkers.

More specifically, the fact-checkers, although under a common objective, have characteristics that distinguish them:

- **FerretScot** (UK): an independent organisation based in Scotland that focuses mainly on issues concerning Scotland and the UK. Although details on the fact-checking techniques used are not specified, the platform is described as a model of journalism that seems to integrate modern methods and advanced technology, thus revealing a mixed human-technology approach ("https://theferret.scot/about-us/").
- **FullFact** (UK): an independent organisation founded in 2010 in the UK with the aim of fact-checking information and news on a wide range of topics. This fact-checker, like its predecessor, has a main focus on news about and from the UK, although it also sporadically ranges over international news. FullFact uses various technologies for fact-checking, such as software developed in-house using models such as BERT (Bidirectional Encoder Representations from Transformers). The use of artificial intelligence technologies certainly helps to improve service coverage, although, as they point out, human verification remains an indispensable process ("https://fullfact.org/ai/about/").
- **PagellaPolitica** (Italy): an independent organisation founded in 2012 with the aim of verifying information and statements by mainly Italian politicians, although it also covers other areas such as the economy. There is no reference to the methodologies used by the organisation, although it is emphasised that there are multiple actors involved in each verification. It is part of the fact-checking organisation Facta ("https://pagellapolitica.it/manifesto").
- **FactaNews** (Italy): an independent organisation that concentrates mainly on news concerning Italy, dealing in particular with news circulating on social networks, with a wide range of topics. The use of tools to identify viral news that will then be verified is mentioned, but the type of technology used is not specified ("https://facta.news/chi-siamo/").

- **AfpFactuel** (France): a service initiated by the Paris-based international news agency Agence France-Presse (AFP), the only one of the fact-checkers we cover provided by a news agency. Although AFP Factuel is a French service, it checks news internationally, leveraging the global network and resources of the news agency it is run by. The fact-checking process combines manual verification by journalists with the use of technological tools, mainly used for image and video analysis, to assess whether or not they are generated by artificial intelligence ("https://factuel.afp.com/comment-nous-travaillons").
- **CheckNewsfr** (France): an organisation founded in 2017 in France, it is part of the same country's newspaper "Libération". Unlike the other fact-checkers analysed, it uses a participative approach, i.e. it invites users to ask questions about news they would like to have verified by professionals. This organisation focuses mainly on news pertaining to the French context, although it does not exclude the checking of international news with global impact. It is not specified whether the fact-checking methodologies used also include the technology approach ("https://www.liberation.fr/checknews/").

The data was collected using the Twitter API. An API, or Application Programming Interface, is a set of rules and protocols that allow two software components to communicate with each other.

In this case, the Twitter API therefore acted as a bridge between the social network's database and external users, allowing public information to be extracted.

The latter can be various and of different types. For the purposes of this study, the data concerns the tweets published by the fact-checking accounts with the relevant interaction and engagement measures, the retweets also with the relevant measures as well as the name of the user who re-posted, the list of users who were following the fact-checking pages at the time when the request for data was made to the API.

The basic information, obtained via the Twitter API, is collected in 3 different datasets for each fact-checker, according to the following scheme:

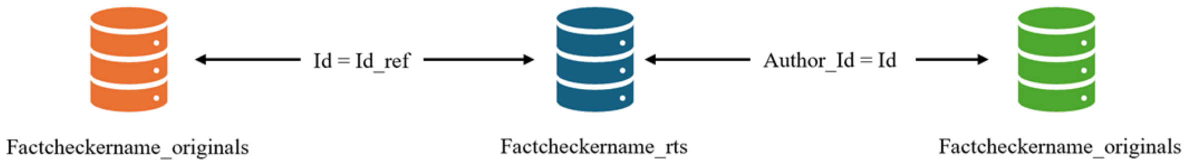

Figure 1: Dataset's Schema

In detail, each dataset contains the following data:

- **Factcheckername_originals**:
    - Tweet ID: unique identifier for each of the tweets
    - Tweet text: textual content of the message
    - Date and Time: date and time the tweet was created
    - Engagement measures: number of likes, number of impressions, number of retweets and number of replies
    - Type of Tweet: i.e. whether the tweet is an original, a retweet or an embedding of another user's tweet in your own in order to comment on it
- **Factcheckername_rts**:
    - Retweet ID: unique identifier for each retweet
    - Tweet text: text of the original retweeted tweet
    - Author ID: unique identifier associated with the user who retweeted the tweet
    - Retweet-related engagement measures: number of likes, number of impressions, number of retweets and number of replies
    - Original Tweet ID: identifier of the original tweet
- **Factcheckername_followers**:
    - Author ID: unique identifier of each follower on the page
    - User name: Twitter username of the follower

	Original Tweets	Tweet Retweeted	Followers
FerretScot	737	5544	32597
FullFact	194	19418	231861
FactaNews	478	394	6813
PagellaPolitica	1201	16490	49340
AfpFactuel	670	13689	164221
CheckNewsfr	611	14832	77147

Table 1: Number of observations for each dataset

In addition to the raw data just described, obtained directly from the Twitter API, other information is obtained through feature engineering techniques, as we will see more specifically later.

Feature engineering is a fundamental process in data analysis that allows the creation or extraction of new features from the raw data, to obtain new information useful for deepening the dataset, but also to provide machine learning models with more data.

The latter point is of considerable importance because it allows the models' effectiveness to be increased by making the process more accurate.

Furthermore, GPT's NLP model was exploited for the extraction of sentiment analysis from the tweet text. The methodology and motivation as to why this was done will be explored later.

The combined approach of feature engineering and the use of external templates for the extraction and creation of new information from the data collected via APIs represents a multidimensional approach that exploits the full potential of the data, both from an exploratory and insight point of view and to feed the statistical models.

To ensure the independence of the analyses for the individual fact-checkers and to foster the consistency of the results between them, the analysis and sequencing of the various processes were applied equally to the different datasets.

This methodological approach ensures that the conclusions this study arrives at are entirely objective, to universally reflect user behaviour and trends on the social media under consideration (Twitter).

This is reinforced by the fact that, as mentioned above, the fact-checkers come from different countries, enriching the generality of the study and its transnational impact.

3.2. Methods

The analysis carried out in this study focuses exclusively on data from fact-checking pages extracted, as described extensively in a previous section, via the Twitter API.

The purpose of the study is to provide insights with the aim of assisting these profiles to improve not so much the content of the tweet, but the manner in which the content is transmitted.

We particularly recommend a strategy that leverages intrinsic characteristics of the tweet such as its length, the presence of emoji, likes, hashtags, and sentiment of the text content.

This approach aims not so much to maximise the engagement of tweets, but more specifically, aims to reach the highest number of users who do not currently follow the fact-checking page under consideration.

Through this analysis, news verification pages can adapt and refine their communication strategies to reach the users most exposed to fake news and, more generally, to disinformation, creating a healthier information environment.

To do this, the main tool used for analysing the large amount of data is the Python programming language.

This has among its main advantages the enormous ecosystem of libraries and frameworks that make it one of the most versatile on the computer landscape.

Libraries are collections of modules and functions that allow actions to be performed in a pre-configured manner without having to write them from scratch.

Of these, those used in this study are:

- **Pandas** and **NumPy**: used for data manipulation
- **Matplotlib** and **Seaborn**: used for data visualisation
- **Scikit-learn** and **Statsmodel**: used for building and implementing machine learning models
- **Scipy** and **Pingouin**: used for statistical techniques and linear algebra
- **Re**: integrated module in Python allowing string manipulation
- **Spacy**, **Gensim** and **NLTK**: used for Natural Language Processing (NLP)

Using these libraries, it was possible to optimise both time and resources, facilitating the handling of complex datasets and large volumes of data, so that advanced statistical methodologies could be applied.

Obviously, Python libraries must be supported by solid theoretical foundations and accurate modelling of reality in order to express their full potential.

Let us now go on to analyse in detail the algorithms and statistical techniques used in the course of the analysis, so that we can then be ready to go over and describe the actual analysis on which this work is based.

3.2.1. NLP and Sentiment analysis

In the context of this research, sentiment analysis was used to enrich the data provided by the Twitter API. To do so, we employed a pre-trained NLP, specifically GPT-4.

We harnessed the capabilities of GPT-4 by crafting specific prompts that directed the model to perform sentiment analysis.

This approach made it possible to classify the news and tweet content published by the fact-checking pages into three different categories: “Positive”, “Neutral” and “Negative”.

With this new information, it was possible to broaden the scope of the analysis by carrying out tests with the aim of understanding the users’ reactions and perception of the sentiment expressed by the message content.

The advantages of using an NLP model in the field of sentiment analysis are manifold and concern in particular the ability to study the emotional complexity of messages, as well as automating the process capable of supporting a very large amount of data, capturing different facets of emotions.

Furthermore, as described in the paper produced by Drus and Khalid (2019), this methodology has, among the characteristics that offer it a great advantage, versatility, finding application in fields ranging from business to social analysis.

3.2.2. Independence Tests

Various statistical tests were used to perform independence and correlation analyses between the different features populating our dataset.

Independence tests are of fundamental importance especially for the data exploration phase, as they allow us to understand and deepen the dynamics within the dataset.

Furthermore, through their application it is possible to identify patterns and trends within the data that may not be apparent, justifying observed results and leading to more informed decisions so as to improve the entire data analysis process.

Also in the modelling phase, identifying strongly correlated variables reduces the problem of multicollinearity, decreasing the complexity of models without sacrificing their effectiveness.

In this study, several correlation tests were carried out. The choice of the type of test depends on the characteristics of the data to be examined, in particular the nature of the features, i.e. whether they consist of qualitative data (categorical) or quantitative data (numerical), and whether they have several classes represented within them or only two (binary variables).

Let us now describe the tests used and the theoretical justifications for their choice:

1) **Chi-square test** -> performed between qualitative variables

The chi-square χ^2 test for the analysis of independence is a statistical tool used extensively to test the possible significant relationship between qualitative (categorical) variables.

After constructing contingency tables, the test helps to understand whether the observed differences between the categories of each variable in the tables are random or are caused by a dependency between the variables (Pandis, 2016).

The chi-square test is based on calculating the observed frequencies in each cell and then comparing them with the expected value if the two variables were completely independent.

The chi-square value obtained from the comparison quantifies the discrepancies between observed and expected frequencies, the higher the value the more unlikely it is that the differences are due to chance. To determine the statistical significance of the differences, the chi-square value is compared with the critical value corresponding to the chosen confidence level, if it is higher it can be concluded that the differences cannot be due to chance, and therefore the variables are not independent ("https://towardsdatascience.com/chi-square-test-with-python-d8ba98117626").

2) **Kruskal-Wallis test** -> performed between qualitative and quantitative variables

The Kruskal-Wallis test offers a non-parametric approach to the analysis of variance for a single classification criterion, being more powerful than the median test (McKight & Najab, 2010).

The test does not require normality of distributions as it is based on ranks rather than their actual values, and it is less sensitive to outliers as it minimises the effect of outliers as it focuses only on the relative order of the data.

The test is effective for testing the null hypothesis that *k* independent groups come from the same population with equal median.

In this study, the Kruskal-Wallis test will be used, as we shall see later, to analyse the effect of a dummy variable on a count variable. The purpose of applying the test to this type of data is to assess the existence of significant differences in the distribution of the count variable between the two groups defined by the dummy variable.

Then, for the tests that resulted in the rejection of the null hypothesis, visualisation along the various percentiles was carried out. In this way, it is possible to identify whether stochastic dominance is present, i.e. whether the curve of one of the two groups is consistently above the other along all percentiles. This provides a more precise and detailed understanding of how one category consistently performs better or worse than the other, on the various levels of the count variable.

3) **Correlation matrix (Spearman)** -> performed between quantitative variables

A correlation matrix is a table showing correlation coefficients between variables.

Each cell in the table contains a correlation coefficient, the value of which can vary between -1, which indicates a perfectly negative correlation, and +1, which indicates a perfectly positive correlation, if the value is 0 this means that no correlation is present.

In our study, we used Spearman's method to calculate the correlation coefficients. This method, unlike Pearson's method which requires the variables to be normally distributed, is a non-parametric test, as it uses the ranks of the variables and not their actual values (Zar, 2005).

The choice to use this method in our study is due to the fact that, being based on ranks, it is more robust in the presence of outliers.

3.2.3. Log-linear Poisson regression

The main model chosen for the analysis is the log-linear Poisson regression.

Poisson regression is a type of generalised linear model that is particularly well suited in cases where it is necessary to model the number of times an event occurs (Coxe et al., 2009).

This regression is used to predict a dependent variable representing a count of events (Hutchinson & Holtman, 2005). The number of these events is usually non-negative and discrete.

Fundamental, is the concept that under certain assumptions, the Poisson can be reduced to a binomial distribution Bin(n,λ/n), with n representing the number of attempts and p, the probability of success, here specified as λ/n.

The assumptions are as follows:

- $n \to \infty$, number of attempts that becomes infinitely large
- $p \to 0$, chance of success, i.e. the probability of the occurrence of a given event, specified here as lambda/n, tending to 0

Under these conditions, considering the parameter λ as constant, the binomial distribution Bin(n,λ/n) converges to a Poisson distribution with parameter λ (expected rate of events).

This mathematical approximation makes the Poisson model particularly effective for count data in which events are rare compared to the number of opportunities in which they may occur.

The log-linear approach of Poisson regression allows a non-linear model to be transformed into a linear one in terms of logarithms, facilitating the interpretation of results.

The formula is the following:

$$\log(\lambda i) = \beta 0 + \beta 1x1i + \beta 2x2i + \ldots + \beta kxki$$

Where:

- λi : expected count of events for the i-th observation
- log(λi) : natural logarithm of λi indicating that the response variable in the model is the logarithm of the event count
- $\beta 0$: intercept that represents the log of expected count of events when all predictors are 0
- $\beta 1,\ldots,\beta k$: coefficients that modified the value of the expected event count. Each coefficient βj corresponds to the effect of the j-th predictor variable.

In this model, each coefficient βj represents the logarithm of a multiplication factor for the event rate when the corresponding variable xij increases by one unit. In mathematical terms, this means that a unit change in xij multiplies the expected event rate by e^(βj).

This makes the coefficients of the Poisson model immediately interpretable in terms of proportional effects.

3.2.4. Holm-Bonferroni correction

The Holm-Bonferroni correction is a method used to control the alpha (type I) error when performing a large number of simultaneous statistical tests. This is because as the number of multiple tests increases, the probability of making such an error increases (Abdi, 2010).

The Holm-Bonferroni correction is based on the Bonferroni correction, which addresses this problem by simply dividing the significance level α by the total number of tests m, and using α/m as a threshold to ascertain the significance of each test (Weisstein, 2004).

As this method is very conservative, the Holm-Bonferroni method was developed from its concepts, with the aim of increasing the statistical power of the tests.

This correction is divided into the following stages:

1. All m p-value of the comparative tests are sorted in ascending order
2. Each p-value is compared, following the order established previously, with a significance threshold that matches its position in the sequence i.

$$Treshold_i = \frac{\alpha}{m-i+1}$$

Where:

- α : level of significance (0.05)
- m : total number of comparative tests
- i : position of a specific p-value in the ordered sequence

3. As soon as a p-value is found that is not statistically significant according to the predetermined threshold, all subsequent values in the sequence will also be considered non-significant.

In conclusion, the Holm-Bonferroni correction is useful in studies with a moderate number of comparative tests where the Bonferroni correction might be too conservative (Sedgwick, 2014).

3.2.5. Keywords analysis

The extraction and determination of keywords is a fundamental aspect of text analysis and natural language processing.

In this context, the data preprocessing phase is essential to ensure that the data follows a structure that conforms to the requirements of the analysis algorithm, which we will see later.

In our study, the preprocessing steps carried out are as follows:

- **Tokenization**: the text is transformed into a list in which the elements are the words from which the tweet is composed.
- **Stopwords removal**: stopwords are words that are frequent in texts but do not bring semantic meaning to the text, such as articles (e.g. “the”, “a”) and conjunctions (e.g. “and”, “but”). The most popular method for removing stopwords is the comparison of predefined, obviously language-specific stopword lists with the text to be processed (Saif et al., 2014). The elimination of these stopwords leaves only the words that influence the meaning of the text, improving and enabling operations such as keyword extraction and text classification.
- **Lemmatization**: the process by which words are reduced to their basic form, taking into account their grammatical role in the text (e.g. conjugations of the verb to be such as “is” or “are” reduced to “be”). The latter is the main difference from stemming, which simply removes prefixes and suffixes without considering the linguistic context (Plisson et al., 2004). Python’s “spacy” library allows the use of a predefined language model based on the language of interest, so that inflected forms of words can be mapped to the corresponding lemmas.
- **Dictionary creation**: fundamental is the creation of the dictionary that maps each unique word to a unique identifier. Using the “gensim” library, not only is the numerical ID assigned to each unique word, but the overall frequency of the word within the corpus is also recorded. In this way, after using more specific patterns and analyses, it is possible to trace the words back to their respective IDs, which facilitates the interpretation and visualisation of the results. The dictionary is also necessary for the transformation of documents into Bag of Words as we shall see in the next section.
- **Transformation in Bag of Words (BoW)**: after creating the dictionary, it is important to transform the documents (tweets) into a numerical form. In our study, each document is transformed into a list of tuples (word_ID, frequency), where each

tuple represents a unique word in the document and its frequency within the document (Qader et al., 2019). The BoW is often the format required, as in our case, by NLP algorithms.

3.2.6. Keyness

The statistical methodology used in this study to establish word relevance is Keyness.

This method is based on the comparison of the frequency of a word in the target corpus with respect to another one, called the reference corpus. Key words will be those that appear in the target corpus with a significantly different frequency than would be expected based on the distribution of words in the reference corpus. The terms that deviate positively will be identifiable for the context of that corpus.

One tool that can be used to calculate keyness is the log-likelihood (Pojanapunya et al., 2018). This measure compares the probability that the frequency of a word in the two corpora is the result of chance against the alternative hypothesis that it is not.

The formula for the log-likelihood G^2 is:

$$G^2 = 2\sum_{i=1}^{k} O_i \, log(\frac{O_i}{E_i})$$

Where:

- O_i represents the observed frequency of the word in each corpus (target and reference)
- E_i represents the expected frequency of the word assuming an identical distribution in both corpus
- k represents the number of groups (corpus) compared, in this case is equal to 2
- i represents the reference corpus

It calculates the expected frequency of tokens for both the target and reference corpus, based on the size of the corpora and the frequencies of the tokens. Afterwards, it is assessed whether the observed frequencies deviate significantly from the expected ones, the greater the deviation, the more unlikely it is that the difference is due to chance.

Finally, the results are sorted based on the log-likelihood values in descending order, with the words with the greatest discrepancy, i.e. the most identifying of the corpus in question, at the top of the list.

As a basic function, which was subsequently adapted according to the specific needs and objectives of this study, we used the function available on Github (https://github.com/mikesuhan/keyness)

3.3. Workflow overview

The objective of this analysis is to be able to determine which tweet characteristics can be modelled and exploited in order to reach as many users as possible who do not yet follow fact-checking pages.

The first step in achieving this is to collect data and extract the tweet characteristics that are to be taken into account.

In order to extract as many of these features as possible, several techniques were used.

First, we downloaded the tweets published by each of the fact-checking pages via Twitter's API. These already contained, for each tweet, engagement measures such as the number of likes, views and retweets.

Afterwards, we obtained, by means of feature extraction techniques, the intrinsic measures concerning the composition of the tweet, such as the number of words, emojis, the presence of links, quotes or hashtags.

These features were also useful in the exploratory phase to compare the communication methods used by each of the fact-checkers.

As described above, another feature of fundamental importance is the sentiment of the tweet, which helps to understand how the emotion conveyed by the tweet can be perceived by the readers. We obtained this information through the use of a pre-trained NLP model.

Finally, by performing the join between the 3 different datasets obtained from the API, i.e. the dataset containing information about the original tweets, the dataset containing information about retweets and the dataset containing the list of users following the fact-checking page at the time the data was downloaded.

By combining all these datasets, we obtained the field that contains the most important information of all, namely, for all retweets, whether the user who retweeted the message is a follower of the fact-checking page or not.

This information represents what will be the target variable of our model. In particular, we are going to build two models, one that has as target value the information on users who have re-shared the tweets and are followers of the page and, vice versa, a second model in which the target value will be the number of users who have re-shared the post but are not followers of the fact-checker.

Our goal is not to use the model to predict an outcome, but instead what we want is to open its black box and go and study how each of the features affects the probability that the user who re-shared the tweet is or is not already a follower of the page.

3.4. Analysis

3.4.1. Creating target variables and sentiment analysis

The first stage, as in all analyses, is the collection of data via the Twitter API.

For each profile of the various fact-checkers, we have three datasets, one containing the original tweets published by the page and the engagement measures attached, one containing the retweeted text, the name of the user who retweeted the post and the engagement measures, and finally, a dataset containing only a list with the names of the users who at the time of downloading the data were following the fact-checking page profile.

By means of the third dataset, we obtained the key information for the entire analysis. It was in fact used, through a join between the different tables, to obtain two new variables, both binary. The first binary variable will take the value 1 if the user who re-shared the tweet follows the fact-checking page and 0 otherwise, the second column, complementary to the first, will be used during the feature augmentation phase.

This field was merged with the retweet table, on which the dataset that will be taken as input by the machine learning models will be based.

The distribution of the target variable for each of the six fact-checkers is as follows:

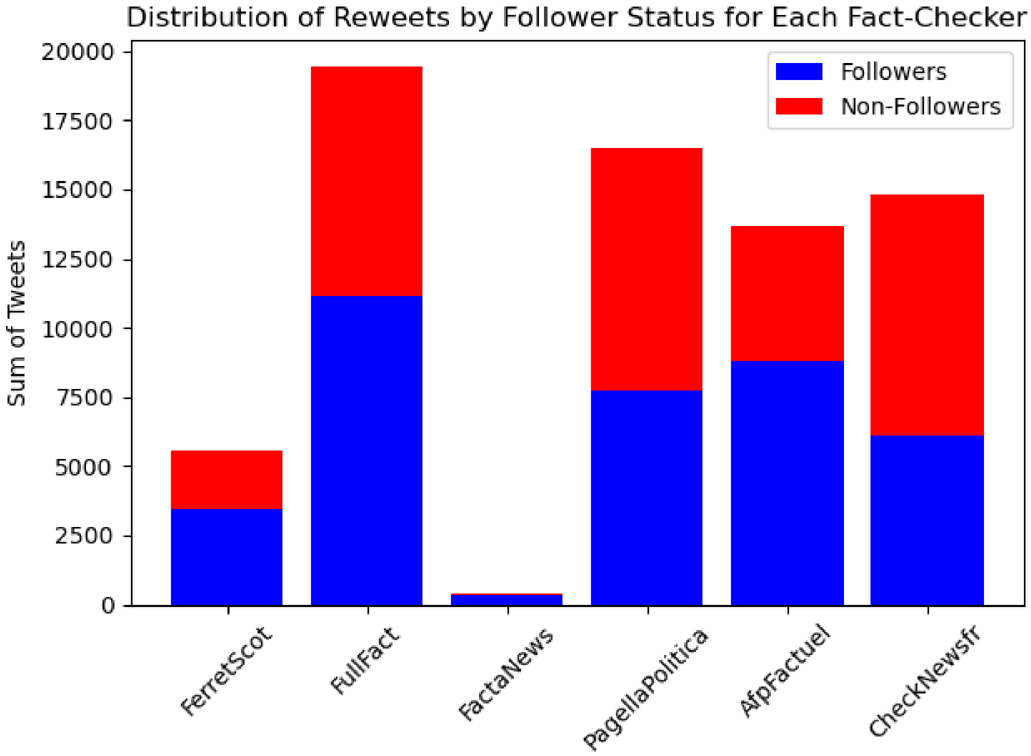

Figure 2: Distribution of retweet by followers status for each fact-checker

Another step, also described above, concerns the sentiment analysis of the original tweets.

In fact, the field of the dataset of the original tweets containing the text of the tweets was given as input to an NLP model, with the aim of obtaining as output a field, associated with each tweet, capable of defining whether the sentiment is “Positive”, “Neutral” or “Negative”.

In the stacked bar chart below, it is possible to examine the number of original tweets, for each of the six fact-checkers, broken down by sentiment:

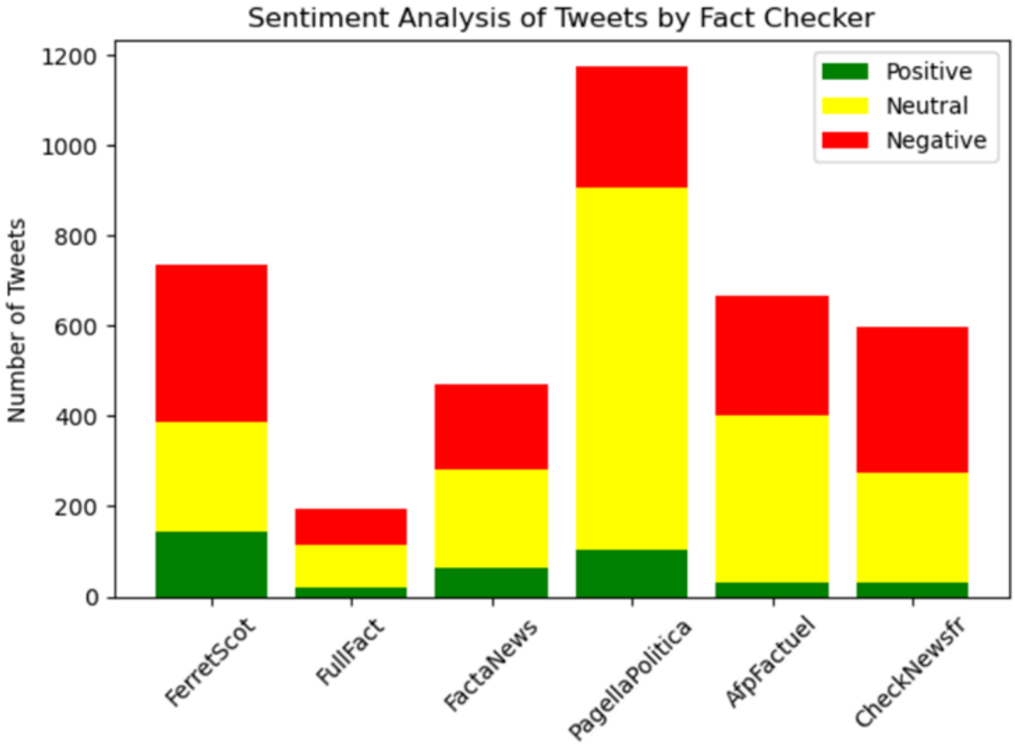

Figure 3: Sentiment analysis of tweets by fact-checker

3.4.2. Data cleaning and Feature augmentation

The purpose of this step is twofold and in both cases of fundamental importance.

On the one hand, one wants to restore the text of each original tweet to its basic form, i.e. without links, hashtags, emojis and any other value "extraneous" to the mere message, to minimise the bias that can be generated by extra-textual elements, in favour of the objectivity and quality of the analyses that will be carried out.

On the other hand, one absolutely does not want to eliminate this information, which can contain key features of tweets in order to reach users who do not yet follow fact-checking pages.

Therefore, for each of the characteristics we will examine below, three steps were performed for each of the tweets published by the page:

1. Search for each of the extra-text objects (emojis, links, hashtags, mentions and quotes) in the tweet

2. Save the information in a new field

3. Remove the objects from the tweet text

The second and third points are the same for all types of objects sought, while the first depends on the characteristics and composition of the objects themselves.

- **Emojis**: each emoji is associated with a unique identifier consisting of sequences of hexadecimal numbers, each of which has a Unicode code point (Davis & Edberg, 2015). This allows emojis to be read and recognised by an operating system.

  😀 = U+1F600 💥 = U+1F4A5 💛 = U+1F49B

  Figure 4: Example of association between Unicode and emojis

  We leveraged this feature to recognise emojis within tweets.

  After defining a pattern that includes a series of Unicode ranges, each representing different types of emojis, we used functions of the “re” module (https://docs.python.org/3/library/re.html), which handles regular expressions, to search for emojis that match the pattern we defined. Finally, a field was added to the dataset to account for the number of emojis identified for each tweet.

- **Links**: in order to identify the presence of links within the tweets, a regex pattern “http\S+” was created, corresponding to each type of sequence beginning with “http”, and then checked for a match with the text of each tweet. If the test is positive, it means that the tweet contains one or more links, otherwise not. This information was also saved in a special field before removing the links from the text of the posts.
- **Number of words**: to count the number of words in each tweet, after removing links and Unicode because they were not recognised as such, simple Python functions were used, and the information was then saved in a new field.
- **Mentions**: the method used to search for mentions is the same as that used for links, except that the content of the string searched for within tweets is only an “@” in this case.

- **Quotes**: the search for quotes within tweets involves a slightly more complex regex pattern than the previous “["\’](.*?)["\’]”. The latter searches for all strings contained within quotation marks ‘‘ or double-quoted “”. The strategy of including both is due to the fact that newspapers, on the basis of custom due to the country in which they are located, may quote in inverted commas or in double quotes. In addition, there is also a condition that in order to be recognised as a quote, the string must contain at least three words. This is because texts within quotes often consist of names, which typically are two or three words long, rather than actual quotes. Therefore, we implemented this method to avoid such scenarios.
- **Hashtags**: the presence of hashtags is checked almost identically to the search for mentions, the only difference being that in this case the “#” character is searched for.
- **Followers/Not followers**: the objective of this last operation is to sum up for each unique identifier of the original tweets, the number of times the tweet is re-shared by followers, and the number of times it is re-shared by non-followers. Then after this operation you will have two additional columns representing the count of one situation or the other.

The dataset resulting from this feature extraction phase will contain the text of the cleaned original tweets, the sentiment of the tweet obtained from the NLP model, the engagement measures downloaded via the Twitter API, the newly obtained features concerning the characteristics of the post and the number of retweets from both followers and non-followers.

3.4.3. Data preprocessing

The dataset, the result of the steps discussed so far, comprises a collection of tweets, each of which is categorised according to several metrics, including the widely discussed categorical “sentiment” and “type” metrics.

The “sentiment” column categorises tweets into three classes “Positive”, “Neutral” and “Negative”, as well as “type” which distinguishes original tweets based on their nature, categorising them as “Original”, “Retweeted” and “Replied_to”.

What has been done is to create dummy variables (Jolly & Gupta, 2021) for each of the classes of the “sentiment” and “type” variables.

By following this methodology, not only will the data be transformed into a more efficient and readable format before being passed into the machine learning model, but the creation of dummy variables allows for independence tests to be performed between the categorical variables and the other variables, in ways that we will explore in more detail in the next section.

Therefore, the categories that observations can take will be converted into a series of binary variables. Each category will be represented by a separate variable that assumes 1 if the tweet belongs to that class or 0 otherwise.

After the stages of feature engineering and preprocessing, the dataset prepared for conducting the independence tests and the subsequent analysis will include the following features:

- **Impression count:** total number of times each tweet has been view
- **Number of emojis:** count of emojis
- **Presence of links:** binary indicator showing whether a tweet contains a link
- **Word count**: count of words
- **Number of mentions:** count of other users mentioned
- **Number of hashtags:** count of hashtags
- **Presence of quotes:** binary indicator showing whether a tweet contains a quote
- **Count of retweets from followers:** count of retweet from followers
- **Count of retweets from non-followers:** count of retweet from non-followers
- **Negative sentiment:** binary indicator showing whether the tweet sentiment is negative
- **Positive sentiment:** binary indicator showing whether the tweet sentiment is positive
- **Quoted:** binary indicator showing whether a tweet is quoting another tweet
- **Replied:** binary indicator showing whether the tweet is a reply to another tweet

The variables "Neutral" and "Original," as we can see, do not appear because they have been set as reference categories for the "sentiment" and "type" variables, respectively.

Before being able to perform independence tests on the dataset, or use it as input for models, a fundamental pre-processing step before running the machine learning algorithms is normalization. Normalization allows us to prevent any one feature from dominating over the others due to the scale on which the data it contains is calibrated; to do this we will transform, for all features, the data into a range [0,1] based on the minimum and maximum values each contains. This process also aids the future interpretation of the output coefficients of the models, which can be compared more directly.

3.4.4. Independence tests

Independence and correlation tests can be considered as a fundamental step in the data exploration phase, although, unlike the previous one, here the aim is to identify the dependence between variables or to quantify the strength and direction of the relationship between them.

The choice of the type of test to be applied depends on the characteristics of the variables one wishes to study. Different combinations of features require different approaches, it is therefore essential to select the most appropriate type of test for each case study.

For the sake of simplicity, only the display of the Kruskal test and correlation matrices for the “FerretScot” fact-checker alone will be shown in the body of the study. The analyses for the other five will be included in the “Appendix” section of the work, available for anyone wishing to verify the information provided in the text.

As for the chi-square tests, on the other hand, which do not include a graphical display, the p-value results, once the due correction has been applied, will be presented in tabular form for ease of reading.

In our specific case, it is important to specify that when tests are carried out that only concern information about the original tweet, such as the correlation matrix and the Kruskal test, the analysis is performed on the dataset relating to the original tweets. However, when the analysis requires the incorporation of the target variable, i.e. the status information of the user who retweeted, it is necessary to conduct these tests on the retweet dataset.

1) **Chi-square test**

   Tests used between qualitative variables, in particular we compare two by two all the categories, transformed into dummy variables in the pre-processing phase, present in the variables “type” and “sentiment”. For each fact-checker, 9 chi square tests are then carried out, the results of which will also be validated using fdr correction, through the “statsmodel” library, with the aim of limiting the risk of committing type I errors (false positives) when multiple statistical comparisons are carried out simultaneously.

For each category of the qualitative variables “sentiment” and “type”, which contained the classes “positive”, “neutral” and “negative” on the one hand and “original”, “retweeted” and “replied_to” on the other hand, respectively, dummies were generated.

Nine tests were carried out for each fact-checker, based on the combinations of each of the “sentiment” variables coupled with each of the “type” variables.

The results that we are going to show below, already have within them the Holm-Bonferroni correction carried out to control and limit the number of false positives that could emerge from the multiple use of the independence tests.

To ensure simplicity and clarity, as it would have been impossible to display all the tests in a single table, we will now focus solely on the chi-squared independence tests for the fact-checker "FerretScot". Data on other accounts will be available in the "Appendix" section.

	Replied	Quoted	Original
Negative	0.001 **	0.049 *	<0.001 ***
Positive	0.014 *	0.037 *	0.005 *
Neutral	0.131 .	0.744 .	0.131 .

Table 2: Chi-square independence test for the fact-checker "FerretScot". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, . >= 0.05.

2) Correlation matrix

Test used between quantitative variables, using Spearman’s method for calculating correlation. The choice of this method of calculation is due to the fact that in the fact-checker datasets there are outliers in some variables, such as viral tweets that have a disproportionate number of impressions compared to the average, and this method is robust in this respect since, as seen above, it is based on ranks and not on actual values.

The results, for the fact-checker considered, are as follows:

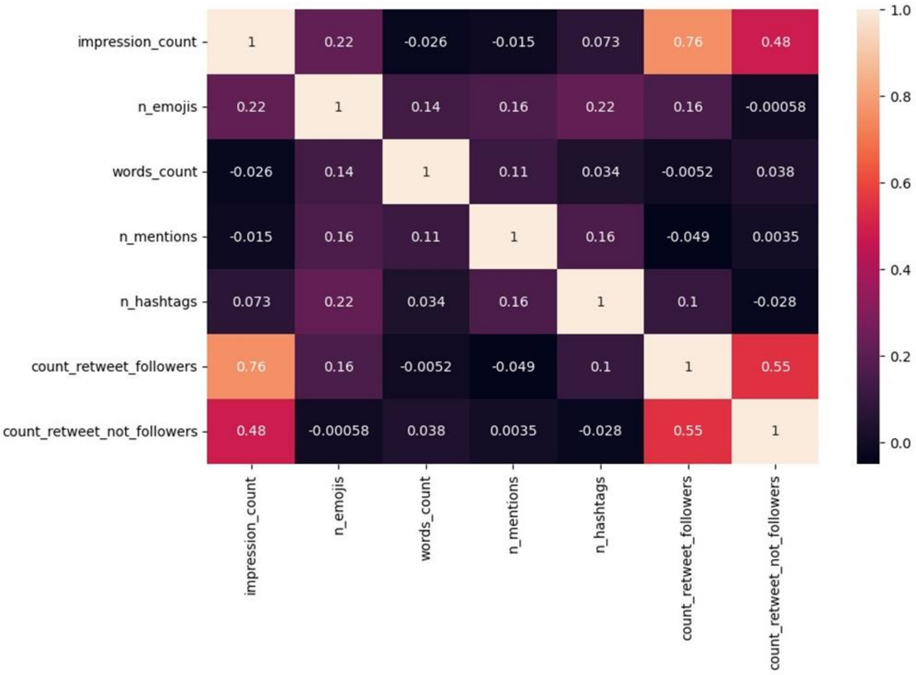

Figure 4: Correlation matrix for the fact-checker "FerretScot". In each cell, the coefficient of correlation is based on Pearson's method.

From this graph we can see that no significant correlations emerge between the quantitative variables, with the exception of that between the variable "impression_count" and "count_retweet_followers". This highlights the fact that the greater the number of times a tweet is seen, the greater the number of times it will be retweeted by users following the page. There is also a relationship, albeit less strong, between impressions and the number of retweets by users who do not follow the page.

3) Kruskal-Wallis test

Test used between the quantitative variable “impression” and each of the dummy variables: “Positive”, “Negative”, “Neutral”, “Original”, “Replied_to”, “Retweeted”, “Presence_links” and “Presence_quotes”. As mentioned above, graphs will be displayed with the aim of verifying the stochastic dominance, carried out by comparing two groups of associated tweets (dummy variable) with respect to the variable “impressions” based on their percentiles, of a single fact-checker.

With regard to the dataset referring to the tweets of the Scottish fact-checker “Ferret_Scot”, we now examine the tests in which the p-value associated with the statistic H is below the chosen significance level of 0.05. In these cases, the null hypothesis is rejected, suggesting that significant differences exist between the dependent variable “impression_count” and the two groups identified by the dummy variable.

	Replied	Quoted	Original	Positive	Negative	Neutral	Presence Links	Presence quotes
Impression	< 0.001 ***	0.009 **	< 0.001 ***	0.333 .	< 0.001 ***	0.013 *	< 0.001 ***	0.030 *

Table 3: Kruskal-Wallis test for the fact-checker "FerretScot". In the cells, the adjusted p-value for each pair of features according to the following logic: $p\text{-value} \le 0.05$, ** $p\text{-value} \le 0.01$, *** $p\text{-value} \le 0.001$, . >= 0.05.*

For each of the cases where the p-value is less than the chosen significance level (0.05) we display the number of impressions for the two groups associated with the dummy variable for each percentile. If one of the two groups has higher impressions for all percentiles then it would signify a case of stochastic dominance.

Again for the sake of simplicity, we will only display 2 cases, one in which stochastic domination is present and one in which it is not.

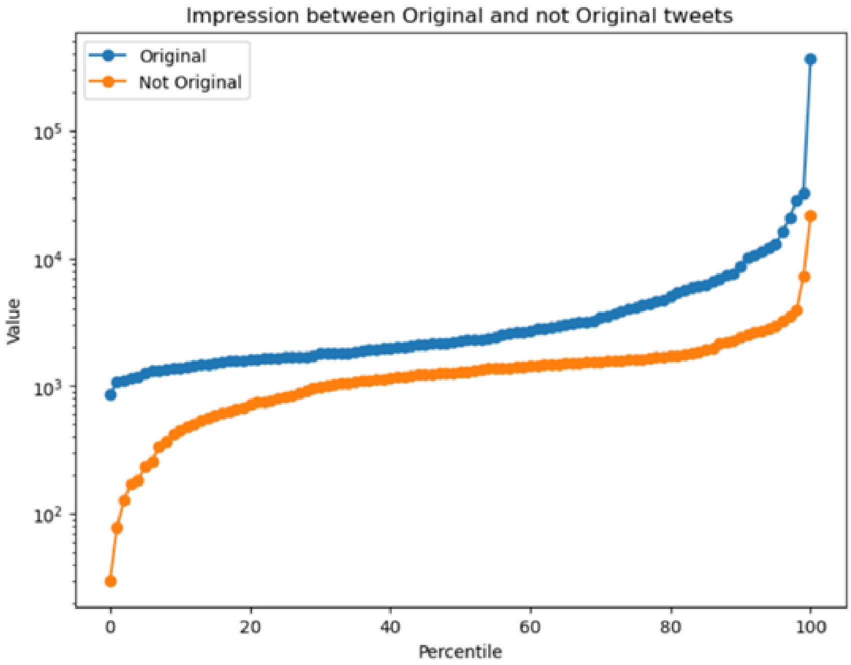

Figure 5: Number of impressions by percentile for two groups, categorized by the "Original" dummy variable (0 and 1) for the "FerretScot" fact-checker.

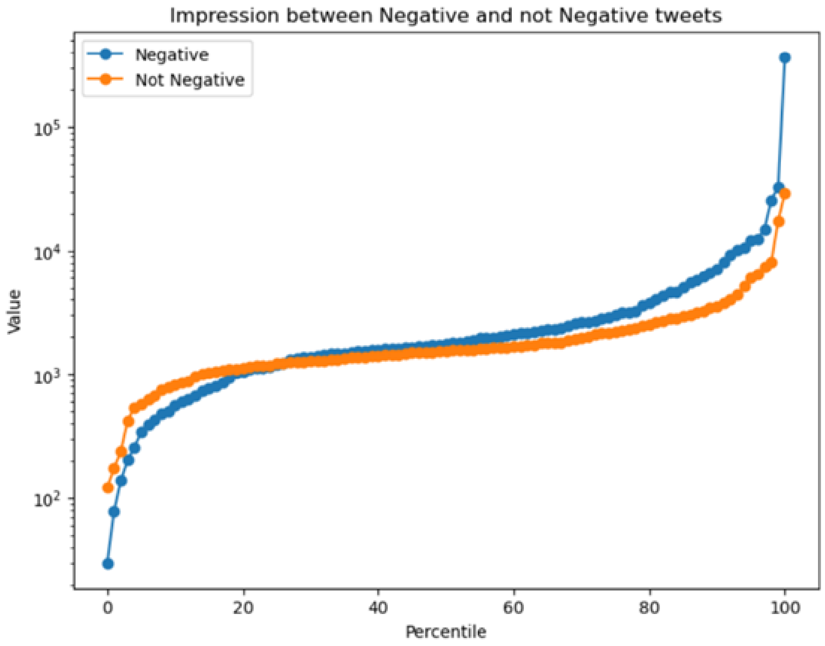

Figure 6: Number of impressions by percentile for two groups, categorized by the "Negative" dummy variable (0 and 1) for the "FerretScot" fact-checker.

What emerges from these visualisations is that in the case of the comparison between the groups of the variable “Original”, stochastic domination is present, whereas in all other cases this cannot be stated.

3.4.5. Modelling

In the context of this study, concerning the analysis of the spread of fact-checker tweets among users who do not follow the checkers’ profiles, a log-linear Poisson regression was implemented to model reality.

This choice proved to be particularly effective in analysing rare events with respect to the number of occasions on which they may occur.

In fact, as described above, the Poisson model assumes that the number of attempts n is infinitely large and that the probability of success for each attempt p tends to zero.

When these conditions are met, the binomial distribution Bin$(n,\lambda/n)$ converges to a Poisson distribution with parameter λ.

This is particularly akin to the characteristics of the data possessed, where the events of interest, i.e. re-shares of tweets, are much less frequent than the views of the tweets themselves. Thus the assumption is fulfilled in that we will associate the number of attempts n with the number of impressions of each tweet, the number of times the event occurs λ with the number of times each tweet is re-shared, and the probability of the event occurring $p = \lambda/n$ tends to zero as required.

To demonstrate that the assumption is respected, for each of the fact-checkers’ datasets, we will now display a graph that contains on the y-axis the count of events, while on the x-axis the ratio of retweets to impressions per tweet will be shown. The visual representation will clearly illustrate that the vast majority of the data points are concentrated near zero, affirming that the probability of the event occurring tends to zero as required.

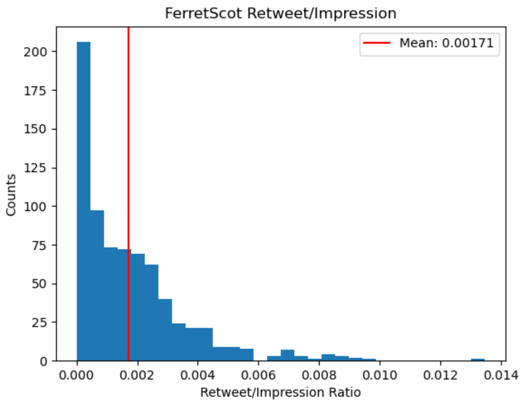

Figure 7: count of events versus the ratio of retweets to impressions per tweet for the "FerretScot" fact-checker dataset.

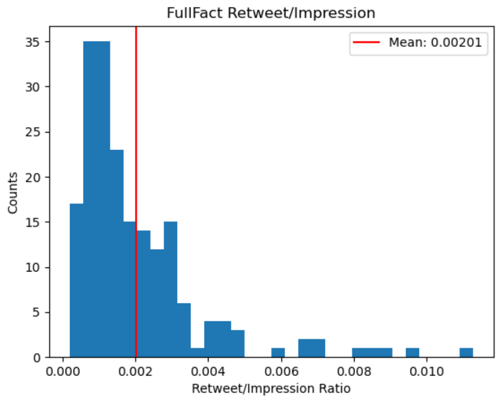

Figure 8: count of events versus the ratio of retweets to impressions per tweet for the "FullFact" fact-checker dataset.

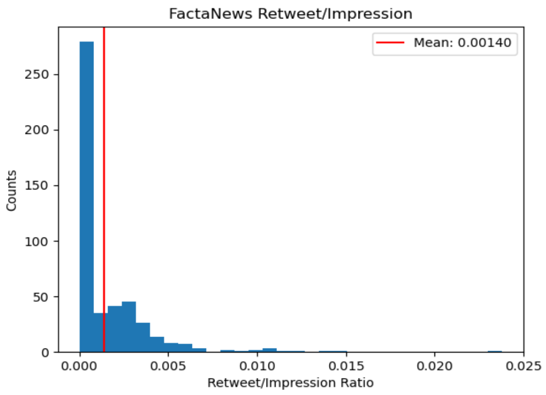

Figure 9: count of events versus the ratio of retweets to impressions per tweet for the "FactaNews" fact-checker dataset

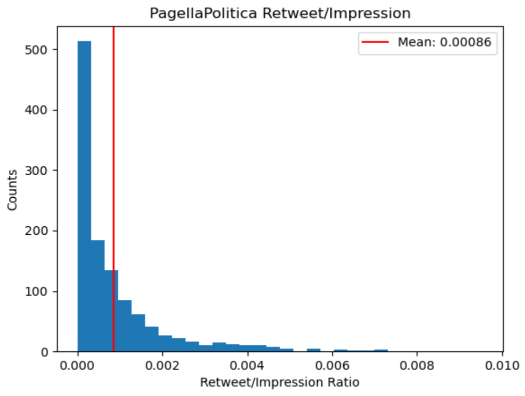

Figure 10: count of events versus the ratio of retweets to impressions per tweet for the "PagellaPolitica" fact-checker dataset

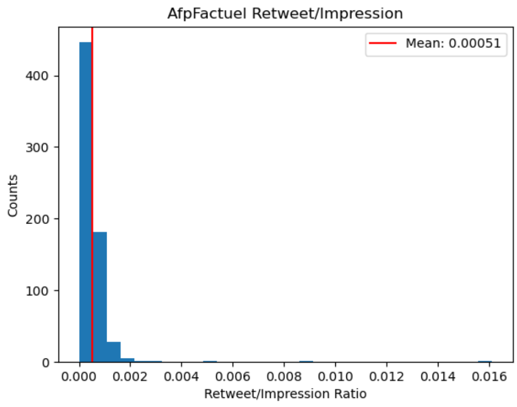

Figure 11: count of events versus the ratio of retweets to impressions per tweet for the "AfpFactuel" fact-checker dataset

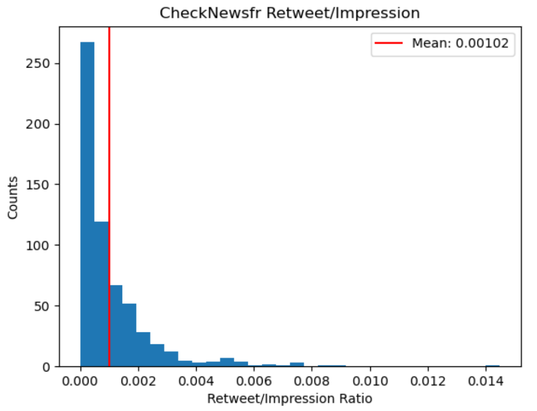

Figure 12: count of events versus the ratio of retweets to impressions per tweet for the "CheckNewsfr" fact-checker dataset

Having defined in the previous points the data that will be taken as input by the model, and the choice of algorithm for modelling reality, we now go on to specify the target variable.

It is of paramount importance to clarify that the target variable of our model is not simply the number of retweets of each tweet, but aims to achieve a more specific goal. The objective is, in fact, to determine the characteristics that a tweet must possess in order not only to maximise the number of users reached, but specifically to reach the highest number of users who do not already follow the page.

To do this, we built two models for each of the six fact-checkers, both with the same input data and both developed on the basis of the log-linear Poisson regression algorithm, but with different target values. In fact, in the first model the target value will be defined as the number of retweets made by those who already follow the page, which we will henceforth define as “followers”, while in the second the target value will correspond to the number of times the tweet is retweeted by users who do not follow the fact-checking page, which we will henceforth define as “non-followers”.

The main goal of this analysis is not to achieve a high level of efficiency and accuracy in predicting the target value on new data, but rather to understand how the characteristics of a tweet influence and affect non-followers.

To this end, we constructed two separate models to analyse not only the impact of these features on the number of retweets by non-followers, but also how these differ from the effects on followers.

This methodological choice also explains why we did not opt for models that are more complex and efficient in terms of prediction, but potentially less interpretable. In fact, for our study, it was necessary to use a model that facilitated the extraction and analysis of the significance and magnitude of each feature on the target variable.

The log-linear Poisson regression proved ideal for the choice of model in this context, as the assumptions and characteristics of the Poisson regression are perfectly suited to the modelling of the phenomena under investigation, while the log-linear approach offers the possibility of deriving a linear model, in terms of logarithm, simplifying the interpretation of the individual coefficients. Furthermore, the log-linear model performs well with outliers, making it particularly robust and reliable for analyzing datasets where extreme values may be present, as in the case here.

In the context of our analysis, we chose not to perform the usual division of the dataset into train and test. This decision is due to the need to maximise the amount of data available to load into the model in order to interpret the effect of the tweet features on the target variable as thoroughly as possible. The standard practice of splitting the train and test set data is essential when the goal is to evaluate the accuracy of a model on previously unseen data, i.e. its ability to generalise what it has learned. However, in our case the priority is to use as much data as possible to enrich the analysis of the significance and magnitude of the tweet features.

For the construction of the model, the Python library "statsmodel" was used, which stands out due to the large number of attributes it offers, which allows us to specifically examine the output of the model from different points of view. This library offers, in fact, advanced features for statistics, such as the in-depth evaluation of the significance of variables, which other technical aspects such as confidence intervals we will examine more specifically later on.

3.4.6. Significance and influence of Tweet characteristics

As a first step, once the model has been constructed, through the output provided by the library, we go on to identify the significant variables. These are the variables that have a statistically significant impact on the dependent variable. We then focus on the coefficients that show an associated p-value lower than the significance level we have established, in this case p-value = 0.05.

However, we cannot rely solely on the level of p-value provided by the model, especially considering the high number of statistical tests that are performed, which could lead to the effect of chance and the generation of false positives, the type I error. To mitigate this risk, and ensure the validity of our conclusions regarding the significance of the variables, we implemented the Holm-Bonferroni correction described above.

Finally, let us turn to the interpretation that allows us to outline the results of our analysis.

After having examined which features have a significant influence on each fact-checker, both when the target value concerns followers and non-followers, let us focus on the type of contribution of each variable.

The first analysis concerns the sign of the coefficient of the different features, to determine whether they positively or negatively influence the target variables. The goal is to identify the features of the tweets that influence the target variables of all fact-checkers equally.

The variables concerning the type of tweet, denoted as “type”, and, sentiment, denoted as “sentiment”, each originally had 3 distinct categories. During the analysis, these categories were transformed into dummies that are passed as input to the model.

From the output of the model, it can be seen that in the case of the classes that were enclosed in “type” within the output of the model, the category “original” will not appear, while in the case of "sentiment", the "neutral" variable will not appear. This is due to the fact that, as we have already said, these categories were chosen as references by the model. The performance of the other categories belonging to the qualitative variables “type” and “sentiment” are not interpreted in isolation, but relative to these reference categories, functioning as a method of comparison to assess the impact of the variables on the target variables.

In the following heatmaps, the first representing the model with retweets from followers as the target value, and the second for the other case study, we can see that each gathers information from the models applied for each fact-checker.

The values in the boxes represent the coefficient of the features, the colour stands for the magnitude and direction, if it tends to red it means that the feature negatively affects the number of retweets while if it is green in a positive way. Empty cells are those in which the feature is not represented by a single element in the dataset or because it is not significant.

The normalization, carried out in the preprocessing phase, does not affect the significance of the individual features, but the absolute value. The value of the coefficient after normalization represents the variation of the target variable, related to the number of retweets, when the starting variable changes its standard deviation. The purpose of normalization is to rescale the features by making the coefficients comparable with each other.

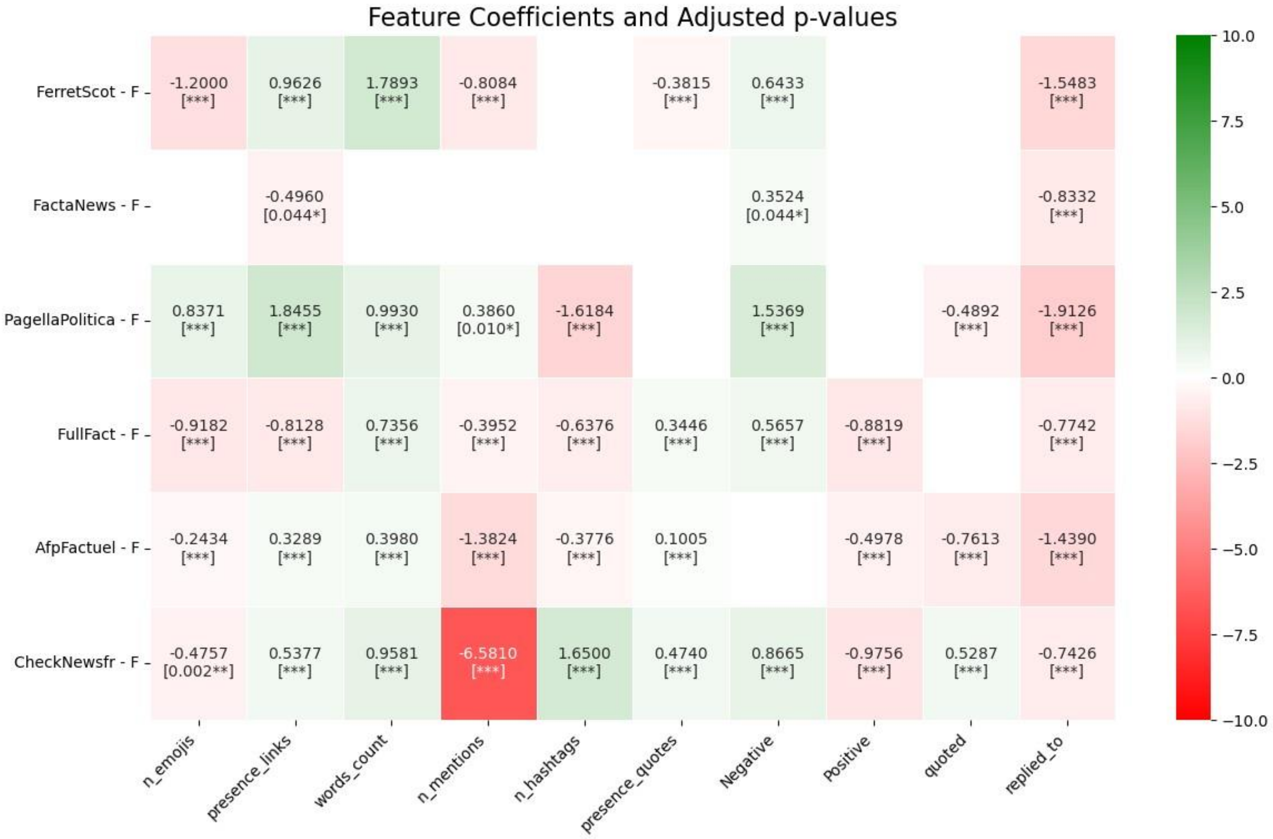

*Figure 13: Heatmap illustrating the model outcomes targeting retweets from followers. Each cell displays the coefficient that determines the color, with negative values in red and positive values in green. Additionally, the significance level of the p-value is indicated in square brackets according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001.*

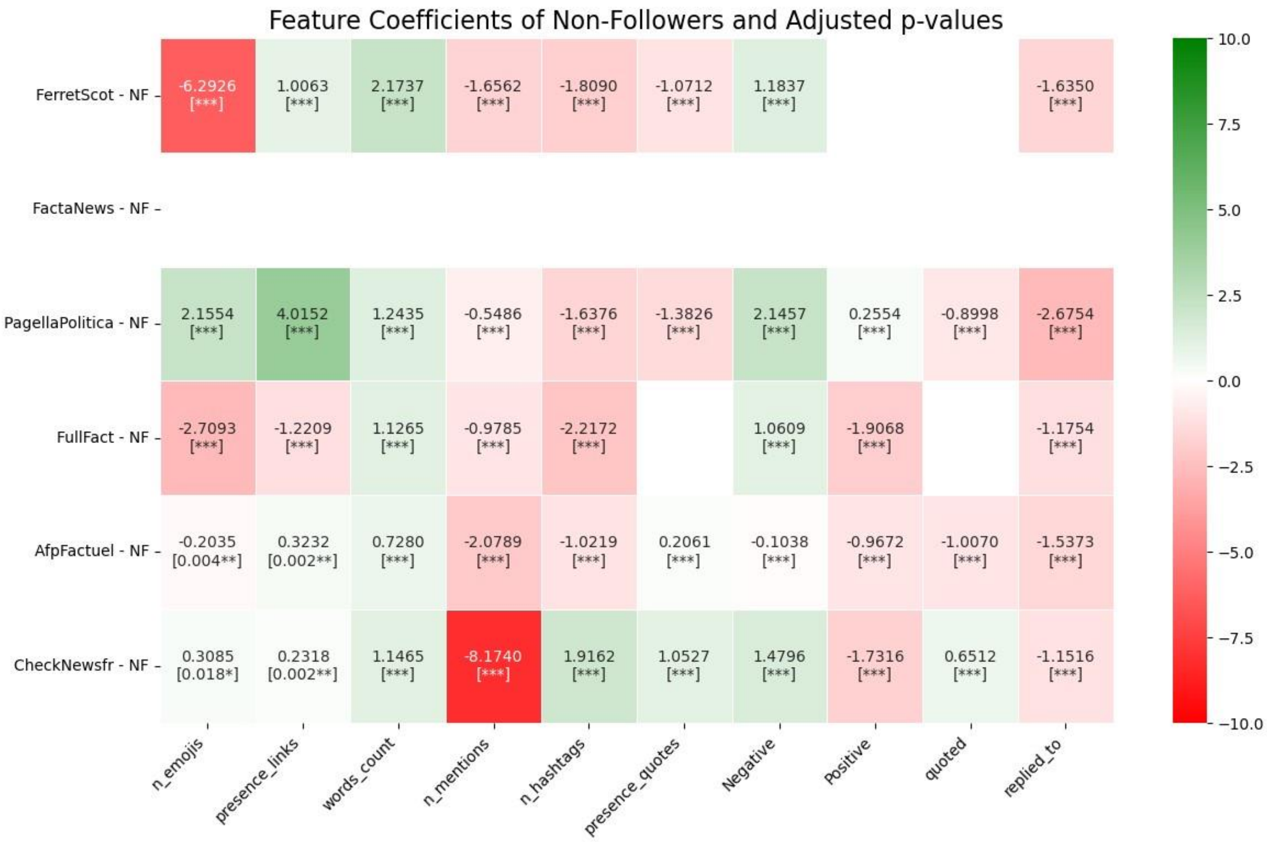

*Figure 14: Heatmap illustrating the model outcomes targeting retweets from non-followers. Each cell displays the coefficient that determines the color, with negative values in red and positive values in green. Additionally, the significance level of the p-value is indicated in square brackets according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001.*

The first thing that can be noticed is that in the case of followers the fact-checker “FactaNews” has only 3 significant variables, whereas in the case of non-followers no variable was found to be significant.

This phenomenon can be clearly explained by the graph “Distribution of Tweets by Followers Status for Each Fact-Checker”. In this visualisation it is, in fact, evident that the number of retweets from the fact-checker “FactaNews” is significantly lower than that of the other pages, in particular the number of retweets from non-followers is such a low number that it is imperceptible in the graph.

The low number of observations means that the variables are not significant because the lack of data prevents a reliable statistical analysis. A small sample size does not provide the necessary generality from a statistical point of view to establish meaningful relationships between the independent and dependent variables.

We can identify characteristics of the Tweets that are consistent, in terms of their impact on the target variables, in both models across all, or nearly all, fact-checkers, demonstrating robustness and consistency. These are:

- **Words_count**: number of words within the Tweet, positive impact on the target variable for both models across all fact-checkers
- **N_mentions**: number of mentions within the Tweet, negative impact on the target variable for both models on all fact-checkers except for "PagellaPolitica" in the case where the target variable is followers.
- **Negative**: type of Tweet with negative sentiment, positive impact on the target variable for both models on all fact-checkers except for "AfpFactuel".
- **Replied_to**: reply to another tweet, negative impact on the target variable for both models on all fact-checkers

3.4.7. Comparison between Follower and Non-Follower

Our focus will now shift to the differences in the impact of features between followers and non-followers. To explore this further, we examine the confidence intervals of the statistically significant features. Let us analyse whether the confidence intervals for the variables related to non-followers show different amplitudes, whether negative or positive, compared to those of followers. If the confidence interval related to non-followers does not overlap with the confidence interval of the same feature of the model related to followers, we conclude that there are significant differences in the way the tweet characteristics influence the two groups.

This allows us to define which tweet attributes are most effective among non-followers, offering insights that may be useful to fact-checking pages when establishing their communication strategy.

Let us now examine the confidence intervals for the two models in each of the six fact-checkers, with the exception of the “FactaNews” fact-checker, as the non-significance of almost all features makes this analysis statistically unreliable.

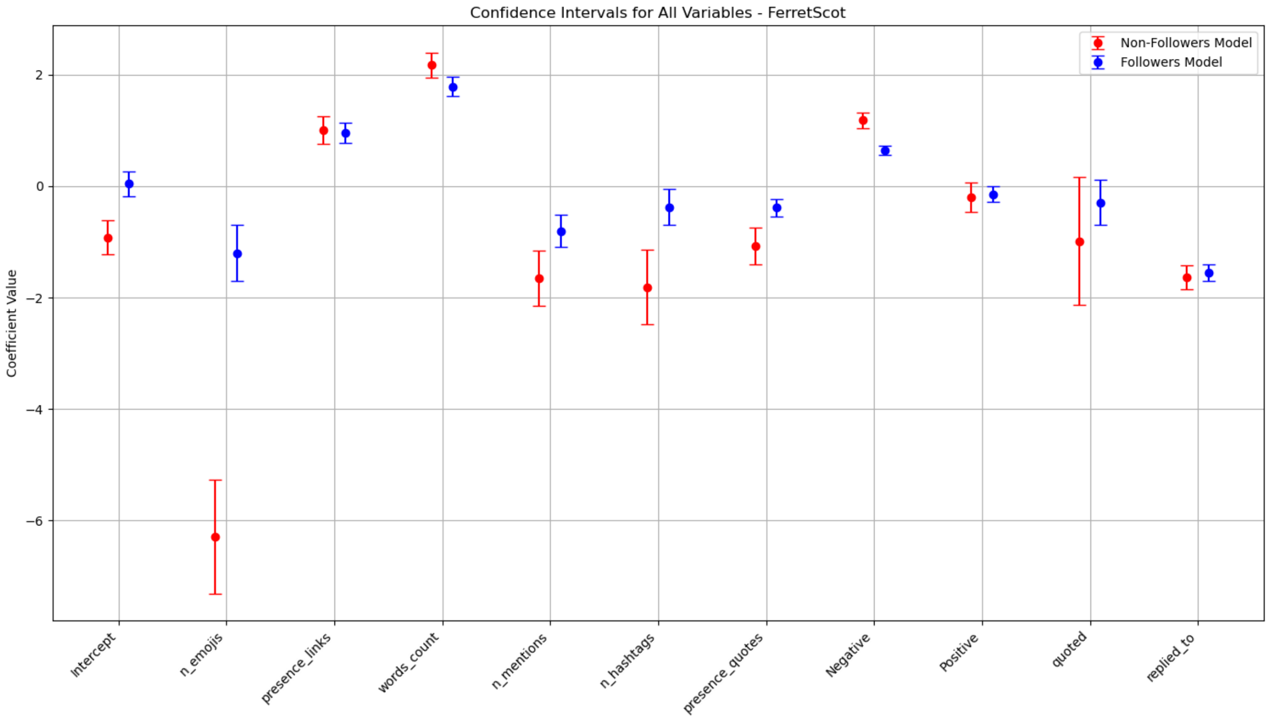

Figure 15: Confidence intervals for each feature, for retweets from both followers and non-followers, regarding the "FerretScot" fact-checker.

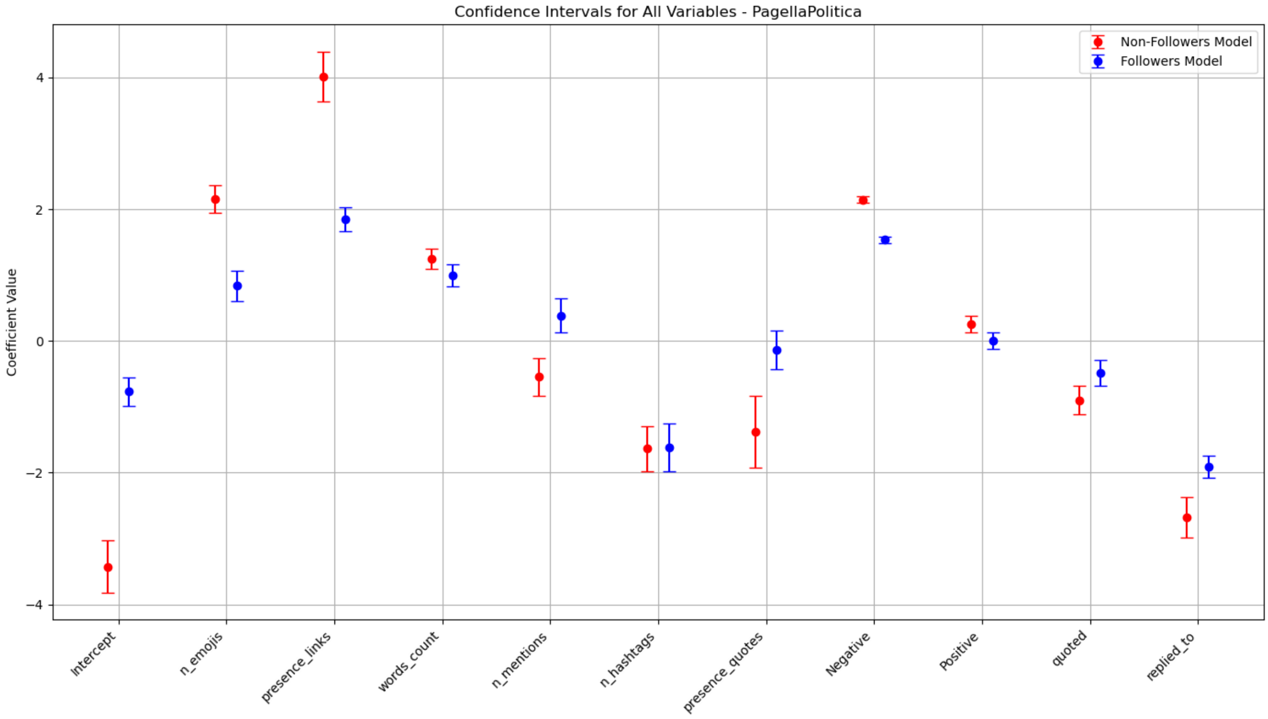

Figure 16: Confidence intervals for each feature, for retweets from both followers and non-followers, regarding the "PagellaPolitica" fact-checker.

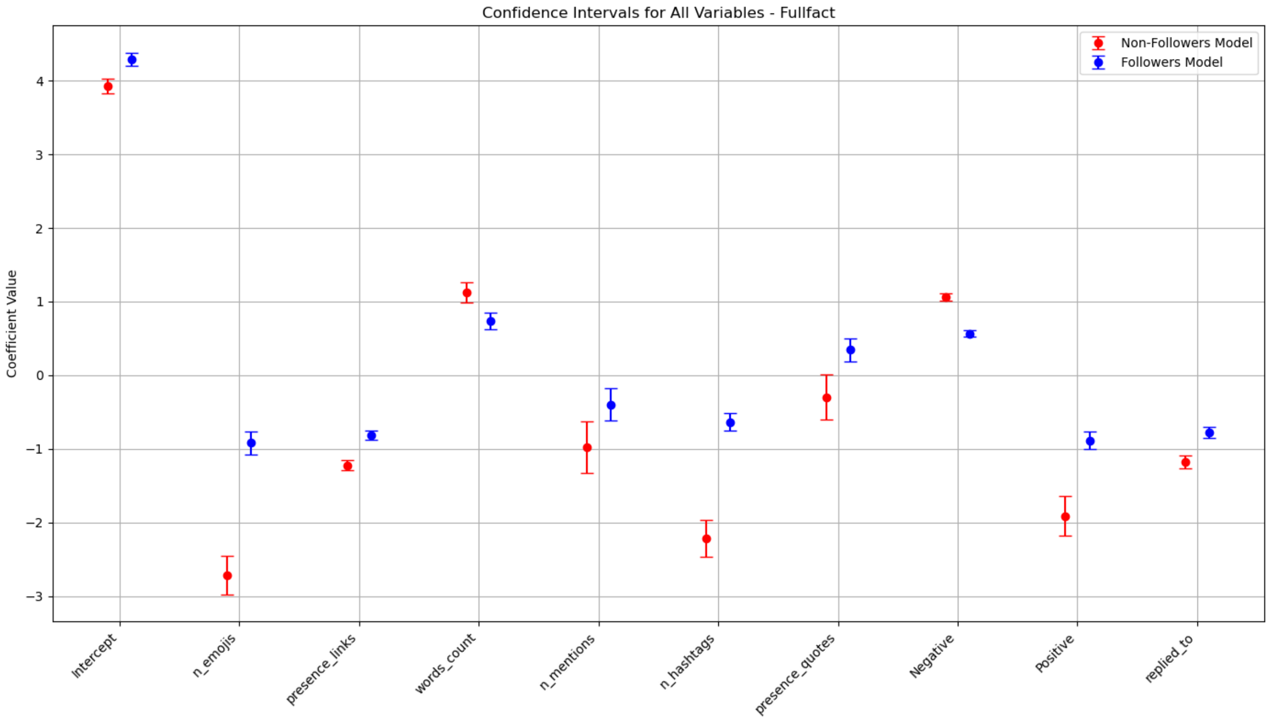

Figure 17: Confidence intervals for each feature, for retweets from both followers and non-followers, regarding the "FullFact" fact-checker.

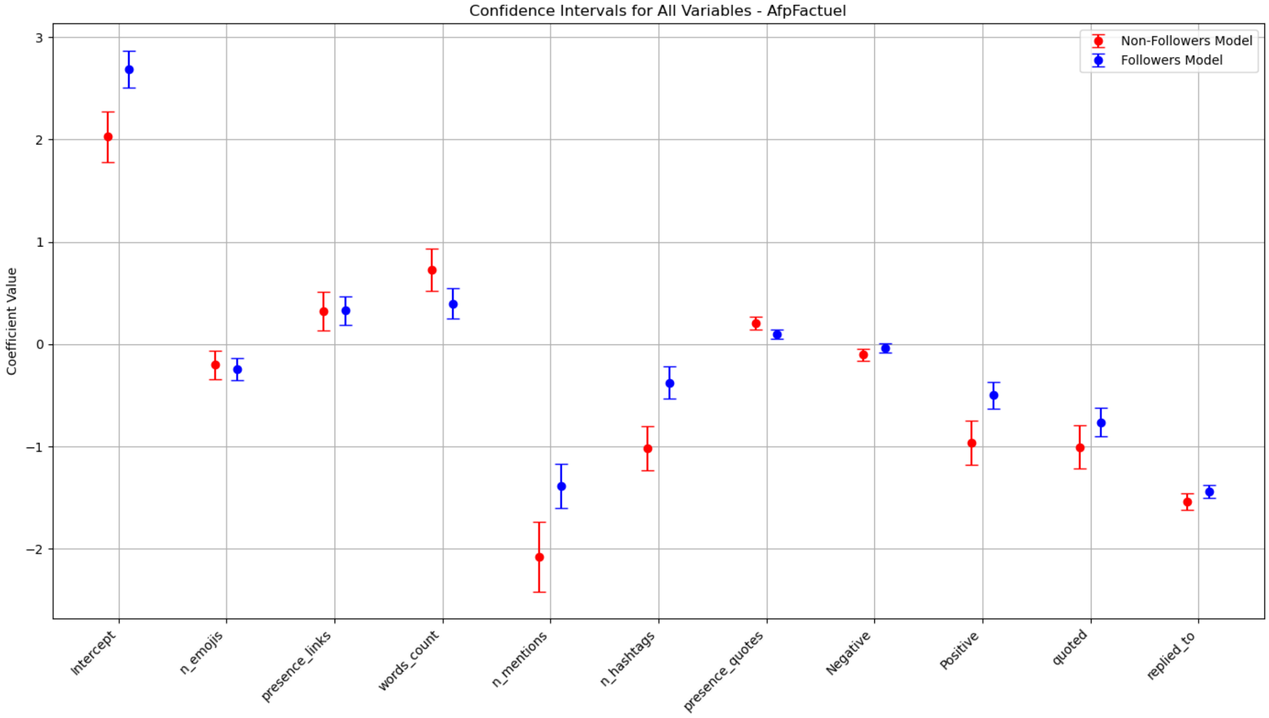

Figure 18: Confidence intervals for each feature, for retweets from both followers and non-followers, regarding the "AfpFactuel" fact-checker.

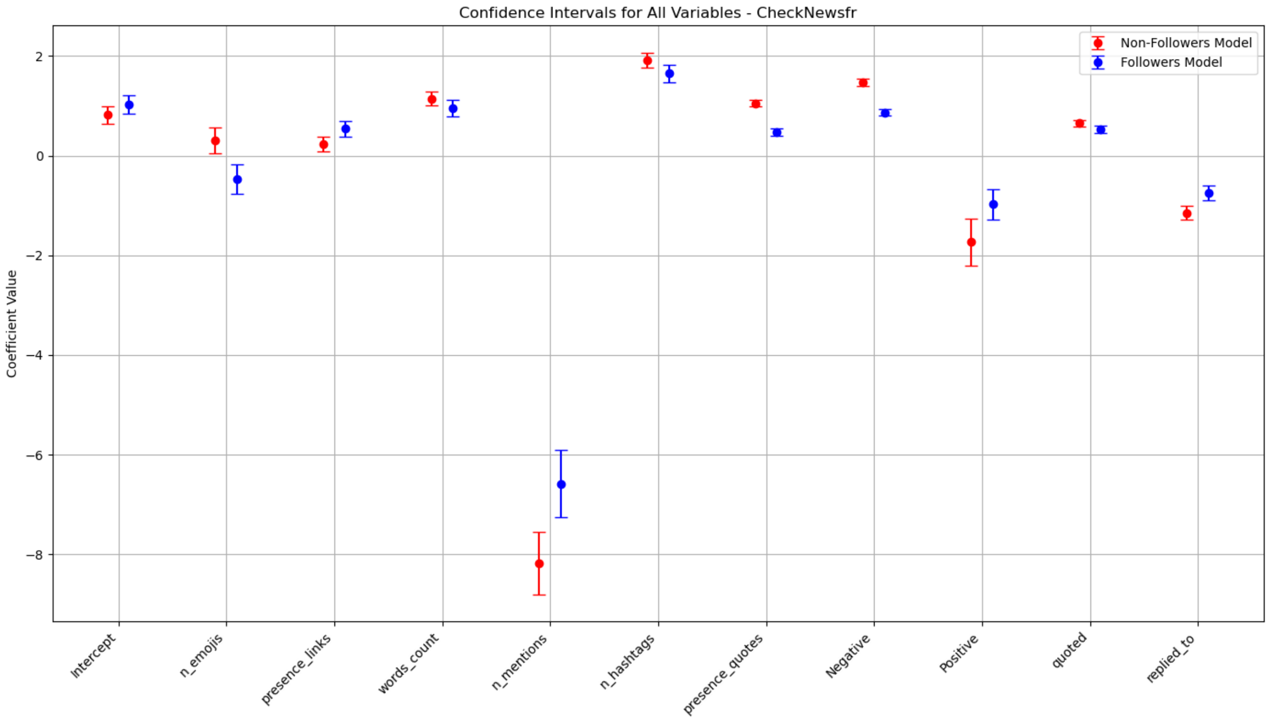

Figure 19: Confidence intervals for each feature, for retweets from both followers and non-followers, regarding the "CheckNewsfr" fact-checker.

Focusing on the 4 features that showed an impact consists in almost all fact-checkers, we propose to test whether the behaviour of these variables differs depending on whether the target variable is related to followers or non-followers.

- **Words_count**: no significant differences emerge between the impact of the feature on followers and non-followers. This lack of differentiation may indicate that content strategies concerning word count are universally effective regardless of the status (followers or non-followers) of the users.
- **N_mentions**: in two fact-checkers there appears to be no significant evidence between the two groups, in the other three seems that an increase in the number of mentions is associated with a stronger increase in the number of retweets among followers than among non-followers.
- **Negative**: for the fact-checker "AfpFactuel" the coefficient in the first model is not significant, indicating that in this case the negative sentiment, compared to the negative ones taken as a reference, does not significantly influence retweets from followers. In all other fact-checkers, not only does sentiment turn out to be a significant variable, but it also shows homogeneous behaviour. In fact, the analysis shows that in all datasets,

the fact that a tweet has negative sentiment is associated with a greater increase in the number of retweets from non-followers than from followers. This suggests that negative tweets may be more effective in capturing the attention of users who do not follow the fact-checker, perhaps due to the stimulating nature of eliciting an emotional reaction.

- **Replied_to**: no significant differences emerged between the impact of the feature on followers and non-followers.

3.4.8. Keyword analysis

The last section of our analysis concerns keyword extraction.

Keyword analysis is crucial for understanding which topics attract the attention of users. Specifically, we are going to examine tweets based on the origin of retweets (from followers or non-followers) or the sentiment expressed (positive, negative or neutral). The aim is, in the specific contexts selected, to identify which keywords emerge significantly.

This analysis is carried out on three different datasets, although the processes are similar. In all three analyses, the pre-processing and formatting steps of the data structure that the NLP model requires are carried out. Hence, the removal of stopwords is carried out, eliminating the terms that do not add semantic value to the text, lemmatization, with the aim of bringing the remaining words back to the basic form or lemma, and, finally, the creation of the dictionary and the transformation of the corpus into numerical data according to the Bag of Words model, which can be managed by statistical algorithms.

The data structured according to the Bag of Words is the input for the Keyness model.

The algorithm based on Keyness was also slightly modified according to the objectives of this study.

When we speak of distance from the mean in the context of log-likelihood, we refer to the discrepancy of the observed data from the mean estimated by the model. However, log-likelihood does not specify in which direction this discrepancy occurs. It is possible for the observed data to be either above or below the mean predicted by the model.

In summary, the log-likelihood tells us how much the observed data deviates from the mean estimated by the model, but does not indicate whether this deviation occurs above or below the mean.

Therefore, unlike the general model described in the section on methodologies, this one will be slightly modified to suit our needs, since we are interested in words that deviate from the mean estimated by the model in the positive.

We have modified the function so that it returns a list sorted by decreasing likelihood with the following data for each element:

- The word
- Its log-likelihood value
- Its frequency in the target corpus
- Its frequency in the reference corpus
- Its relative frequency in the target corpus (absolute frequency in the target corpus / number of words in the target corpus)
- Its relative frequency in the target corpus (frequency in the reference corpus / number of words in the reference corpus).

From these characteristics, we applied a filter such that only keywords for which the relative frequency in the target corpus is greater than or equal to the relative frequency in the reference corpus are displayed, in order to prevent negative log-likelihood discrepancy from the mean.

We chose a very small significance level, p-value = 0.01, to minimise the possibility that the significance of the tests is due to chance. If the log-likelihood value of a word exceeds 6.63 (https://www.gla.ac.uk/media/Media_580807_smxx.pdf) we can conclude with 99% certainty that the presence of the keyword in the target corpus is statistically more significant than in the reference corpus.

I. Non-followers' keywords

The objective of this point is to identify the most important keywords in the tweets that received above-average retweets from non-followers, in order to understand the most significant topics that attract the interest of non-followers.

After calculating the average number of retweets from non-followers, the dataset will be divided into two, on the one hand, all tweets with above average number of retweets from non-followers, on the other hand, tweets with below average number of retweets from non-followers.

Let's now visualize, in descending order by the associated log-likelihood, the keywords corresponding to one of the fact-checkers, in this case, FerretScot.

In the visualization, keywords related to news topics will be highlighted in yellow, while those related to promotional tweets, where the fact-checking page advertises itself, will be highlighted in green.

Fact-checker	**Keywords with number of retweet from NF > Average [LL]**	**Keywords with number of retweet from NF < Average [LL]**
FerretScot	Nuclear [23.637]	Newslett [9.721]
	Hundr [14.182]	Free [8.911]
	New [11.865]	Access [7.021]
	Multin [10.961]	
	Seeker [8.775]	
	Death [8.178]	
	Polic [7.955]	
	Scottish [7.623]	
	Ban [7.069]	

Table 4: Keywords with significance at least 0.01 in the 'FerretScot' dataset. Two comparisons are made: the first between a corpus containing tweets with a number of retweets from non-followers above the average compared to those below the average, and the second the opposite. For each cell, the associated Log-Likelihood for each keyword can be seen in square brackets

The keyword analysis shows that tweets with a higher than average number of retweets from non-followers tend to contain significant keywords related to current topics. In contrast, tweets with a lower than average number of retweets predominantly contain keywords associated with promotional content. These tweets are mainly focused on self-promotion of the fact-checking page and seem to engage less with non-followers, indicating that promotional material may be less effective in generating widespread interest than tweets focused on news topics.

II. Keywords negative

The objective of this point is to identify the most important keywords in tweets with “Negative” sentiment. The choice of this category is due to the fact that consistent results were achieved for this feature in the modelling phase with log-linear Poisson regression, as we will see in the next section.

The aim is then to compare a corpus containing only tweets with negative sentiment with the corpus containing non-negative tweets (positive and neutral), in order to identify the most important keywords for this type of sentiment.

Let us now display, as in the previous case, the significant keywords regarding the FerretScot fact-checker. As before, in yellow are the keywords referring to topical issues, and in green those of self-promotion.

Fact-Checker	**Negative VS Others [LL]**	**Others VS Negative [LL]**
FerretScot	Accus [28.674]	Newslett [50.199]
	Concern [23.488]	Fact [41.9]
	Campaign [23.488]	Subscrib [37.649]
	Group [23.233]	Journal [36.255]
	Warn [17.917]	Ferret [33.959]
	Seeker [17.917]	Free [30.893]

	Fish [17.1]	Sign [27.888]
	Fail [15.16]	Also [26.685]
	Exploit [15.16]	Date [26.447]
	Hotel [15.16]	Sake [24.601]

Table 5: Keywords with significance at least 0.01 in the 'FerretScot' dataset. Two comparisons are made: the first between a corpus containing negative tweets those not negative, and the second the opposite. For each cell, the associated Log-Likelihood for each keyword can be seen in square brackets

We can clearly see that, as we expected, tweets with negative sentiment, compared to those with positive and neutral sentiment, have among their main keywords almost all topics concerning current news. On the other hand, for tweets with non-negative sentiment (positive and neutral), the keywords with the greatest impact all concern the self-promotion of the fact-checker.

III. Absolute frequency

Finally, for exploratory purposes, the absolute frequency of each word within each fact-checker will be calculated. The absolute frequency indicates the total number of times a certain word appears in the dataset.

Through this step, we can gain an overview of the publication patterns and the topics these pages focus on.

This information concerns, considering all the tweets published by the fact-checkers in their entirety, the amount of self-promotion carried out, the main topics and sectors covered and the predilection of the fact-checking page for national or international issues.

Let's now take a look at two fact-checkers, “PagellaPolitica” and “AfpFactuel” analyzing their editorial line through the keywords with the highest absolute frequency.

Fact-checker	Keywords	Absolute frequency
PagellaPolitica	Ital	239
	Govern	227
	Stat	216
	Legg	212
	Part	184
	Melon	166
	Polit	157
	Cos	152
	Ministr	149
	Anno	117
AfpFactuel	Afp	525
	Expliqu	156
	Ce	144
	Imag	102
	Plus	98
	Video	97
	Pouvoir	82
	Vaccin	78
	Faux	73
	Depuis	69

Table 6: Top 10 keywords by absolute frequency for the fact-checkers PagellaPolitica and AfpFactuel.

The analysis shows that the two fact-checkers have markedly different approaches: Pagella Politica focuses on national issues, using keywords specifically related to their own country, while AFP Factuel takes a more international approach, with very few references to national issues. This distinct keyword orientation highlights how the two bodies pursue opposite strategies in the context of fact-checking.

4. Discussion

In the context of the spread of disinformation on social media, where fake news and deepfakes are the main vehicles of dissemination, this study investigated the impact of the characteristics of tweets on user engagement in general, and then of users who do not currently follow the fact-checking page specifically.

This need responds directly to the vulnerability of users, many of whom, as noted by Statista (https://www.statista.com/topics/3251/fake-news/#editorsPicks), do not feel confident in their ability to independently identify fake news. Only 23% of Americans feel confident in being able to recognise such content, while 38.2%, again according to the same source, admit to having mistakenly shared fake news themselves at least once.

At the beginning of this work, we highlighted the need for fact-checking pages to not only sift through as much content and information as possible, but also expand their reach to maximise the impact of their work, only then does it truly gain value.

The goal we set ourselves when writing this study from the very beginning was to build a tool, based on data about fact-checkers and not general data, that could be used by fact-checkers to expand their reach with the aim of reaching the users and people who are most vulnerable to fake news, those who do not follow fact-checking pages.

This study can be regarded as a solid basis for the understanding of user behaviour and engagement dynamics between fact-checkers and users on social networks. One of its aims is to prepaire the way for future developments that can further enhance the impact of fact-checking strategies on social network.

Among these, potential developments include:

- Temporal analysis of engagement: development of a new analysis that also uses as key information when the user started following the fact-checking page, whether before or after the page published a post, or whether before or after the user retweeted the post. This direction of research is currently limited by restrictions imposed by the Twitter API.
- Improved sentiment analysis: In order to improve the objectivity of sentiment analysis, it is important to include alternative methodologies in the classification of different tweet categories with the aim of improving both the accuracy and reliability of sentiment evaluations.

In our research, we opted for a categorization of sentiment into 3 classes “Positive”, “Negative”, “Neutral”, in order to facilitate the most comprehensive analysis possible. However, defining the “Neutral” category proves to be difficult as the boundaries between this and the other categories are very blurred, which is why identifying sentiment objectively and unambiguously for this category is a challenge.

5. Conclusion

This study explored the interaction between the characteristics of tweets and user engagement on social media, focusing on users who do not follow fact-checking pages, who are considered the most vulnerable to misinformation.

We considered 6 fact-checking pages from three countries with the aim of arriving at conclusions that could be consistent and equal for all fact-checking pages, so that we could generalize the results.

First, we carried out the necessary data engineering phase, to expand the dataset by using external models (GPT4), and data preparation, to define a data format and structure suitable for statistical testing and modelling.

After that, an exploratory phase using independence tests and statistical tests allowed us to investigate dependency relationships and the strength of interactions between variables of different kinds.

Chi-square tests for independence were applied with the aim of analysing the relationships between the dummy-transformed qualitative variables, examining in particular the cathegories “Sentiment” and “Type”. Nine tests were performed for each fact-checker, with Holm-Bonferroni correction.

In order to study the relationship between the quantitative variables, we visualised correlation matrices using Spearman's correlation method due to its robustness to the presence of outliers, such as viral tweets that skew the normal distribution.

The results showed that in all fact-checkers there is a significant correlation between the number of views of the tweet and the number of retweets of the same tweet by both followers and non-followers.

The Kruskal-Wallis test was, on the other hand, used to assess the difference between the variable quantifying the number of impressions and the dummy variables representing the type of tweet and sentiment. Only in a few specific cases did significant differences emerge between certain types of tweets, also confirmed by the presence of stochastic dominance with respect to the impression variable.

Following this exploratory test, we moved on to the modelling phase.

First of all, we verified, by means of special visualizations, that the fundamental assumption of the log-linear Poisson regression, our chosen methodology for modeling the phenomenon, was respected in all fact-checkers, we used our pre-processed dataset as input.

We constructed two models, one focused on studying the characteristics of tweets in relation to the number of retweets made by followers, and one in relation to the number of retweets made by non-followers.

The first analysis is the one concerning the coefficients of the significant variables, which allowed us to check which characteristics of the tweets had a significant positive or negative impact on the number of retweets.

The first aspect we noticed is that the "FactaNews" fact-checker has almost all non-significant variables. This is due to the fact that the number of retweets for this fact-checker is significantly lower than for the other pages, especially those from non-followers. Such a small sample size does not, in fact, provide the necessary significance to be able to generalise the relationships between the dependent and independent variables from a statistical point of view.

The most consistent tweet characteristics that emerged were the number of words in the tweet and tweets originating as replies with comments to other tweets. The former has a positive impact on the target variable in all fact-checkers for both models, while the latter characteristic has a negative impact.

Furthermore, tweets containing mentions, which have a negative impact on almost all fact-checkers, and tweets with negative sentiment, which, on the other hand, almost always have a positive impact, showed consistent results on 5 out of 6 fact-checkers.

Then, to achieve what is the specific purpose of this analysis, we went on to examine the differences in the impact of tweet characteristics between followers and non-followers.

To do this, we analysed the confidence intervals of statistically significant features. If the confidence intervals of a certain feature between the model for followers and the model for non-followers do not overlap, this means that there is a significant difference in the way and/or impact with which the feature influences the two groups.

Starting from the features of the tweets that showed a significant and consistent impact between the different fact-checkers based on their coefficients, we came to what is the major conclusion of this study.

Negative tweets are those that show significant differences between the impact on retweets from followers and non-followers in almost all cases, in 5 out of 6 fact-checkers.

Furthermore, in all fact-checkers, with the exception of “AfpFactuel”, tweets with negative sentiment always have a more positive confidence interval in the case of the estimate for non-followers.

This means that tweets with negative sentiment impact the number of retweets positively, and especially positively impact the number of retweets from non-followers compared to those from followers.

Only in “AfpFactuel”, the variable concerning the negative sentiment of tweets, is not significant in the case of retweets from followers.

It is important to note that the significance is closely linked to the category chosen as reference. In our case, for the purposes of interpreting the variables and our objectives, “Neutral” was chosen as the reference. However, it should be pointed out that if we had chosen the category “Positive” as the reference, since the confidence intervals between “Negative” and “Positive” are always disjointed and those of “Negative” always larger, the tweets with a negative characteristic would always have been significant. The only exception in which the confidence intervals of “Negative” are not greater than those of “Positive” is FactaNews, but this cannot be taken into account since, as stated above, almost no variable is significant given the very low number of re-shared tweets.

6. Bibliography

Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of research design*, *1*(8), 1-8.

Agarwal, M., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2023). Audio-visual face reenactment. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 5178-5187).

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. Science advances, 7(36), eabf4393.

Bachmann, I., & Valenzuela, S. (2023). Studying the Downstream Effects of Fact-Checking on Social Media: Experiments on Correction Formats, Belief Accuracy, and Media Trust. *Social Media+ Society*, *9*(2), 20563051231179694.

Bansal, G., & Joshi, M. L. DEEPFAKE: A SYSTEMATIC REVIEW.

BuzzFeed/YouTube. (2018, April 17). You won't believe what Obama says in this video! Retrieved June9, 2021, from https://www.youtube.com/watch?v=cQ54GDm1eL0

Cazzamatta, R., & Santos, A. (2023). Checking verifications during the 2022 Brazilian run-off election: How fact-checking organizations exposed falsehoods and contributed to the accuracy of the public debate. *Journalism*, 14648849231196080.

Chauhan, R., Kansal, I., Popli, R., Kumar, R., & Sharma, A. (2024). Current State of Deepfake Detection and Generation: A Review. In NaN (No. NaN, pp. NaN-NaN). Bentham Science Publishers Ltd.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. Journal of personality assessment, 91(2), 121-136.

Dafonte-Gómez, A., Míguez-González, M. I., & Ramahí-García, D. (2022). Fact-checkers on social networks: analysis of their presence and content distribution channels. *Communication & society*, *35*(3), 73-89.

Davis, M., & Edberg, P. (2015). Unicode emoji. Unicode Technical Standard, 51.

Demuyakor, J., & Opata, E. M. (2022). Fake news on social media: predicting which media format influences fake news most on facebook. *Journal of Intelligent Communication*, *2*(1).

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes?. The International Journal of Press/Politics, 26(1), 69-91.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854.

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science, 161, 707-714.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29.

Engler, A. (2019). Fighting deepfakes when detection fails.

Graber, D. A. (1990). Seeing is remembering: How visuals contribute to learning from television news. Journal of communication.

Graves, D. (2018). Understanding the promise and limits of automated fact-checking. Reuters Institute for the Study of Journalism.

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, *10*, 178-206.

Hao, Karen. "Even the best AI for spotting fake news is still terrible." MIT Technology Review, 03-out-2018 (2018).

Hassan, N., Li, C., & Tremayne, M. (2015, October). Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 1835-1838).

Hutchinson, M. K., & Holtman, M. C. (2005). Analysis of count data using Poisson regression. Research in nursing & health, 28(5), 408-418.

Jo, Y., & Park, J. (2019). Sc-fegan: Face editing generative adversarial network with user's sketch and color. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1745-1753).

Jolly, S., & Gupta, N. (2021). Understanding and implementing machine learning models with dummy variables with low variance. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1 (pp. 477-487). Springer Singapore.

Khan, I. R., Aisha, S., Kumar, D., & Mufti, T. (2023). A systematic review on deepfake technology. Proceedings of Data Analytics and Management: ICDAM 2022, 669-685.

Khan, R., Hazela, B., Singh, S., & Asthana, P. (2022, October). Deepfakes: A New Era of Misinformation. In International Conference on Computing, Communications, and Cyber-Security (pp. 897-908). Singapore: Springer Nature Singapore.

Kornish, L. J. (2006). Technology choice and timing with positive network effects. European Journal of Operational Research, 173(1), 268-282.

Lauer, L., & Graves, L. (2024). How to grow a transnational field: A network analysis of the global fact-checking movement. *New Media & Society*, 14614448241227856.

Lim, C. (2018). Checking how fact-checkers check. Research & Politics, 5(3), 2053168018786848.

Liz-Lopez, H., Keita, M., Taleb-Ahmed, A., Hadid, A., Huertas-Tato, J., & Camacho, D. (2024). Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Information Fusion*, *103*, 102103.

López-Marcos, C., & Vicente-Fernández, P. (2021). Fact Checkers facing fake news and disinformation in the digital age: A comparative analysis between Spain and United Kingdom. Publications, 9(3), 36.

Mania, K. (2024). Legal protection of revenge and deepfake porn victims in the European Union: Findings from a comparative legal study. Trauma, Violence, & Abuse, 25(1), 117-129.

McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. The corsini encyclopedia of psychology, 1-1.

Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2020). Regulating deepfakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice, 15(1), 24-31.

Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2021). “Fake news” is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, *65*(2), 180-212.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., ... & Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. arXiv preprint arXiv:2103.07769.

Nieweglowska, M., Stellato, C., & Sloman, S. A. (2023). Deepfakes: Vehicles for Radicalization, Not Persuasion. Current Directions in Psychological Science, 32(3), 236-241.

Nirkin, Y., Keller, Y., & Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 7184-7193).

Pancer, E., & Poole, M. (2016). The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 US presidential nominees' tweets. Social Influence, 11(4), 259-270.

Pandis, N. (2016). The chi-square test. American journal of orthodontics and dentofacial orthopedics, 150(5), 898-899.

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., ... & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. IEEE Access.

Plisson, J., Lavrac, N., & Mladenic, D. (2004, October). A rule based approach to word lemmatization. In Proceedings of IS (Vol. 3, pp. 83-86). Sn.

Pojanapunya, P., & Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Linguistic Theory, 14(1), 133-167.

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications, and challenges. In 2019 international engineering conference (IEC) (pp. 200-204). IEEE.

Robertson, C. T., Mourão, R. R., & Thorson, E. (2020). Who uses fact-checking sites? The impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior. The International Journal of Press/Politics, 25(2), 217-237.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Sedgwick, P. (2014). Multiple hypothesis testing and Bonferroni's correction. Bmj, 349.

Singer, J. B. (2021). Border patrol: The rise and role of fact-checkers and their challenge to journalists' normative boundaries. *Journalism*, *22*(8), 1929-1946.

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. European Journal of Cognitive Psychology, 18(6), 813-847.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. Plos one, 18(10), e0291668.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society, 6(1), 2056305120903408.

Wang, T. L. (2020). Does Fake News Matter to Election Outcomes?: The Case Study of Taiwan’s 2018 Local Elections. *Asian Journal for Public Opinion Research*, *8*(2), 67-1.

Weisstein, E. W. (2004). Bonferroni correction. https://mathworld. wolfram. com/.

Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022, June). Limits and possibilities for “ethical ai” in open source: A study of deepfakes. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2035-2046).

Zar, J. H. (2005). Spearman rank correlation. Encyclopedia of biostatistics, 7.

Zendran, M., & Rusiecki, A. (2021). Swapping face images with generative neural networks for deepfake technology–experimental study. Procedia computer science, 192, 834-8.

7. Webliography

https://docs.python.org/3/library/re.html

https://facta.news/chi-siamo/

https://factuel.afp.com/comment-nous-travaillons

https://fullfact.org/ai/about/

https://github.com/mikesuhan/keyness

https://pagellapolitica.it/manifesto

https://pagellapolitica.it/progetto

https://theferret.scot/about-us/

https://towardsdatascience.com/chi-square-test-with-python-d8ba98117626

https://www.bbc.com/news/technology-49961089

https://www.gla.ac.uk/media/Media_580807_smxx.pdf

https://www.liberation.fr/checknews/

https://www.statista.com/topics/3251/fake-news/#editorsPicks

https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

8. Appendix

	Replied	Quoted	Original
Negative	0.037 *	0.152 .	0.062 .
Positive	0.015 *	0.553 .	0.015 *
Neutral	0.731 .	0.062 .	0.985 .

*Table 5: Chi-square independence test for the fact-checker "FactaNews". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, . >= 0.05*

	Replied	Quoted	Original
Negative	< 0.001 ***	0.150 .	< 0.01 ***
Positive	0.009 **	0.948 .	0.009 **
Neutral	0.001 **	0.190 .	0.003 **

*Table 6: Chi-square independence test for the fact-checker "PagellaPolitica". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, . >= 0.05*

	Replied	Quoted	Original
Negative	0.327 .	0.229 .	0.229 .
Positive	0.327 .	0.417 .	0.412 .
Neutral	0.229 .	0.327 .	0.229 .

*Table 7: Chi-square independence test for the fact-checker "AfpFactuel". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, . >= 0.05*

	Replied	Quoted	Original
Negative	< 0.001 ***	0.043 *	< 0.001 ***
Positive	< 0.001 ***	0.944 .	< 0.001 ***
Neutral	0.006 **	0.043 *	0.111 .

*Table 8: Chi-square independence test for the fact-checker "CheckNewsfr". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001, . >= 0.05.*

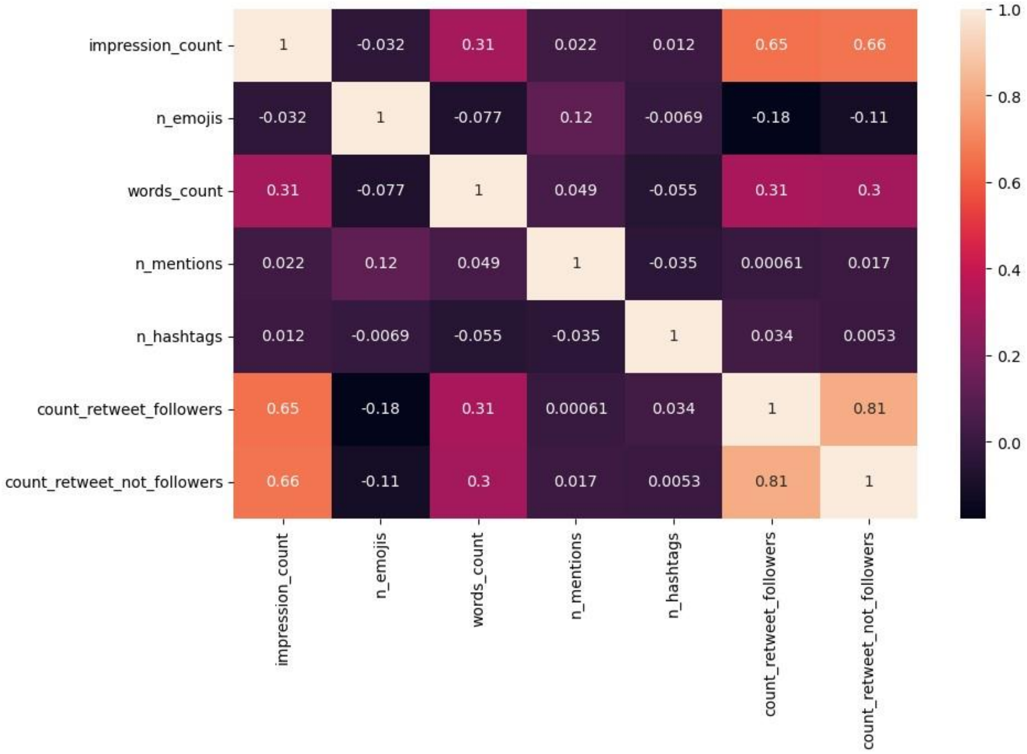

Figure 15: Correlation matrix for the fact-checker "FactaNews". In each cell, the coefficient of correlation is based on Pearson's method.

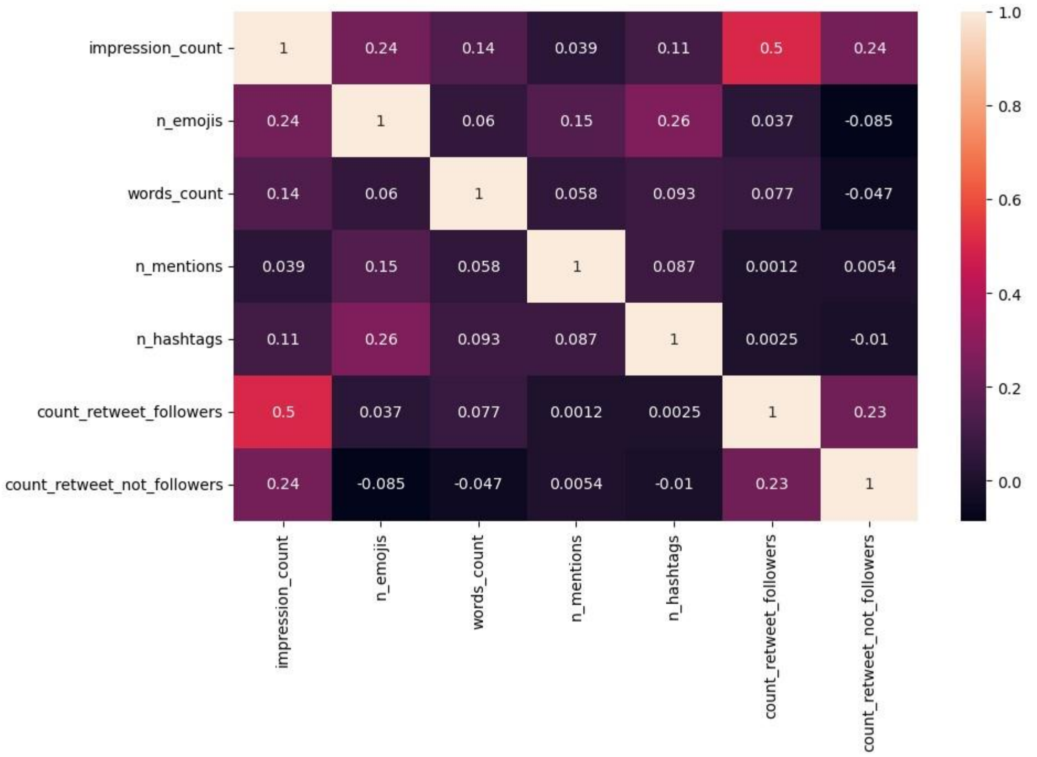

Figure 16: Correlation matrix for the fact-checker "FactaNews". In each cell, the coefficient of correlation is based on Pearson's method.

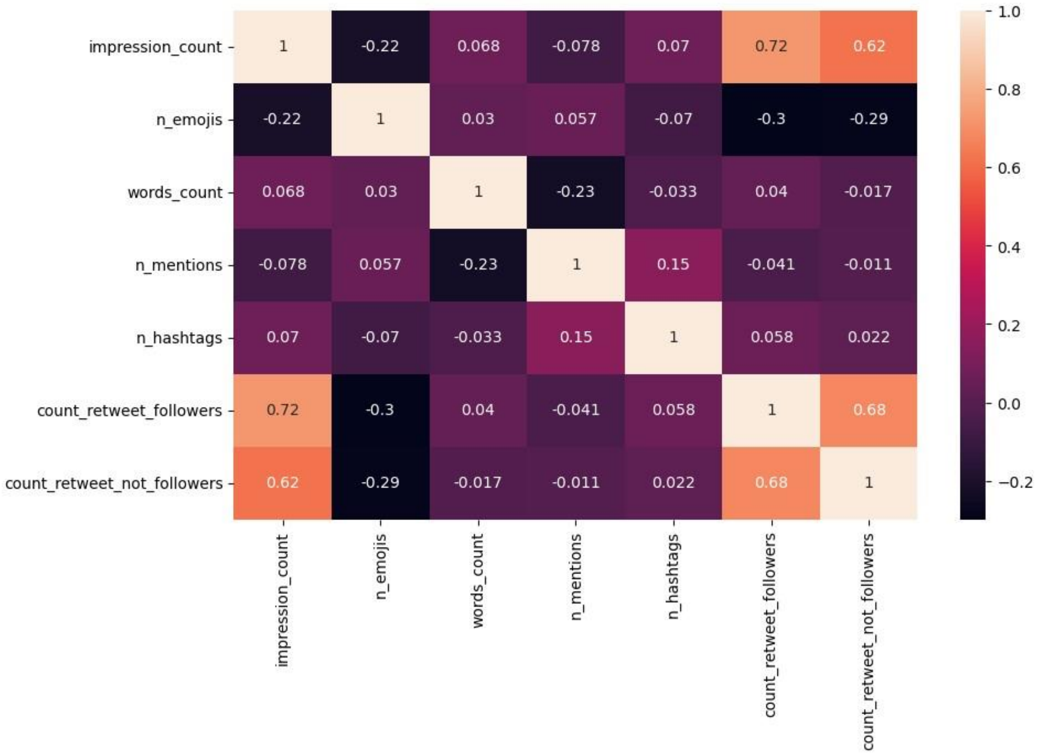

Figure 17: Correlation matrix for the fact-checker "PagellaPolitica". In each cell, the coefficient of correlation is based on Pearson's method.

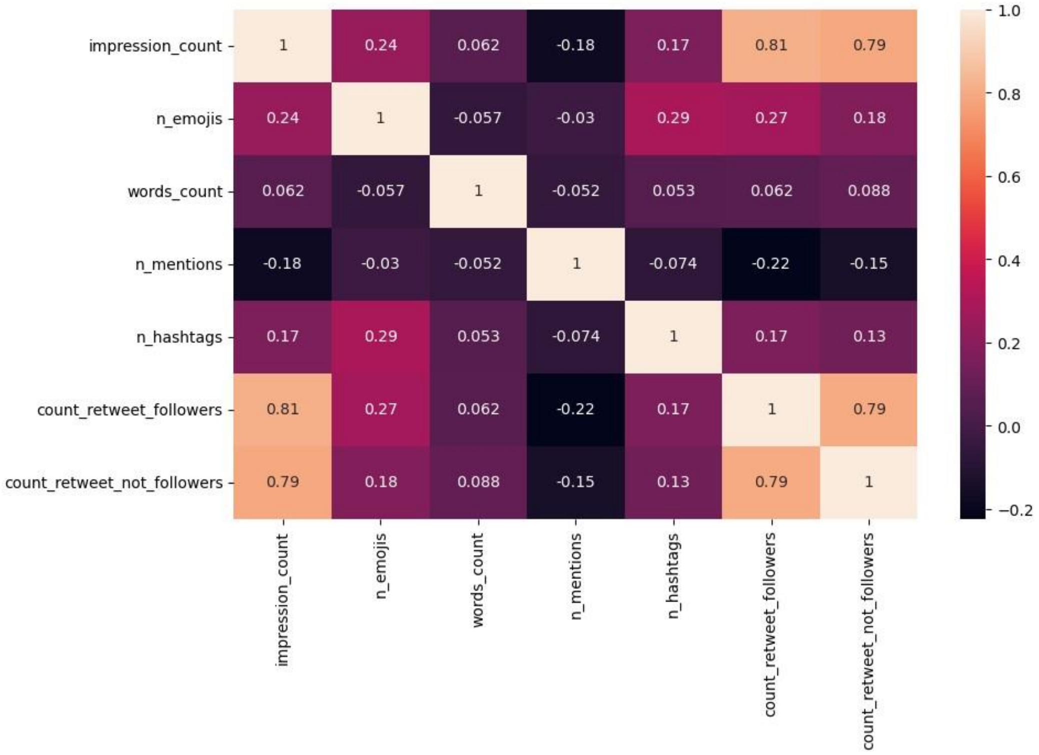

Figure 28: Correlation matrix for the fact-checker "AfpFactuel". In each cell, the coefficient of correlation is based on Pearson's method.

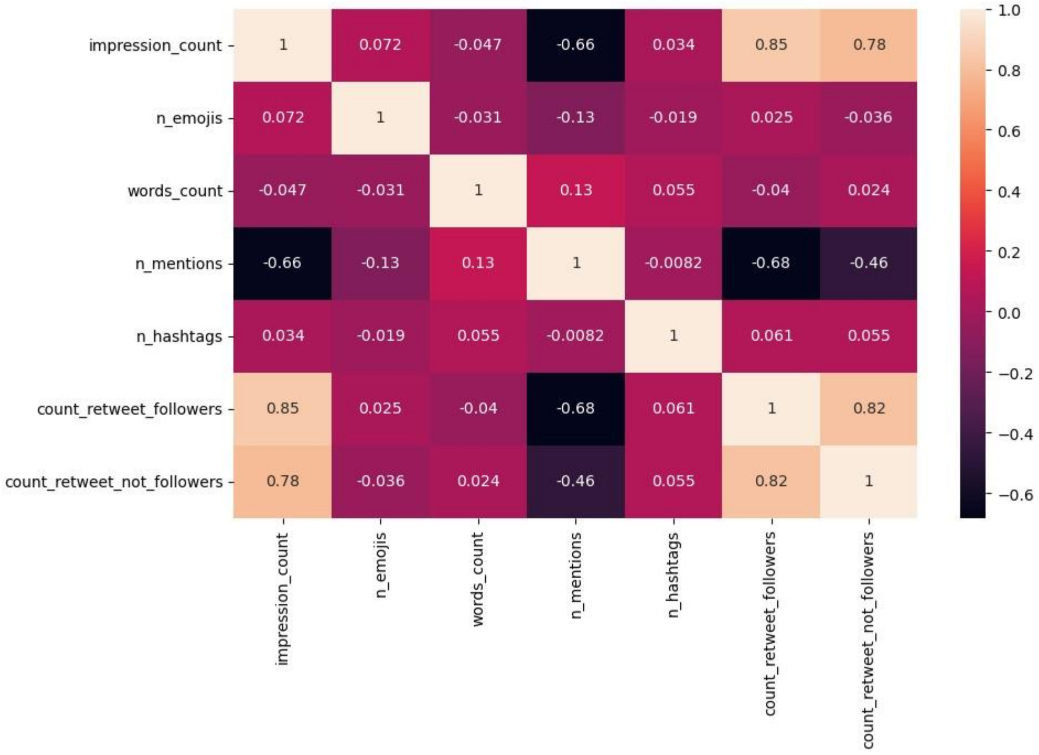

Figure 19: Correlation matrix for the fact-checker "AfpFactuel". In each cell, the coefficient of correlation is based on Pearson's method.

	Original	Positive	Negative	Neutral	Presence Links	Presence quotes
Impression	< 0.001 ***	0.059 .	0.132 .	0.776 .	0.120 .	0.776 .

Table 9: Kruskal-Wallis test for the fact-checker "FullFact". In the cells, the adjusted p-value for each pair of features according to the following logic: p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001*

	Replied	Quoted	Original	Positive	Negative	Neutral	Presence Links	Presence quotes
Impression	< 0.001 ***	0.051 .	< 0.001 ***	0.001 **	< 0.001 ***	0.228 .	0.013 *	0.897 .

*Table 10: Kruskal-Wallis test for the fact-checker "FactaNews". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001*

	Replied	Quoted	Original	Positive	Negative	Neutral	Presence Links	Presence quotes
Impression	< 0.001 ***	0.007 **	< 0.001 ***	0.005 **	< 0.001 ***	< 0.001 ***	< 0.001 ***	0.921 .

*Table 11: Kruskal-Wallis test for the fact-checker "PagellaPolitica". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001*

	Replied	Quoted	Original	Positive	Negative	Neutral	Presence Links	Presence quotes
Impression	< 0.001 ***	0.737 .	< 0.001 ***	0.225 .	0.001 **	0.008 **	< 0.001 ***	0.986 .

*Table 12: Kruskal-Wallis test for the fact-checker "AfpFactuel". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001*

	Replied	Quoted	Original	Positive	Negative	Neutral	Presence Links	Presence quotes
Impression	< 0.001 ***	< 0.001 ***	< 0.001 ***	< 0.001 ***	< 0.001 ***	0.001 **	< 0.001 ***	0.045 *

*Table 13: Kruskal-Wallis test for the fact-checker "CheckNewsfr". In the cells, the adjusted p-value for each pair of features according to the following logic: * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001*