Exploring the Relationship between News Reliability and Violent Comments in Digital Media

Beatriz Botella-Gil[1], Alba Bonet-Jover[1], Robiert Sepúlveda-Torres[1], Patricio Martínez-Barco[1] and Estela Saquete[1]

[1] *Department of Software and Computing Systems, University of Alicante, Spain*

Abstract

Natural Language Processing (NLP) has become an essential tool for the automatic detection of violent language and unreliable information. The misuse of Information and Communication Technologies (ICTs) fosters the generation of disinformation and digital violence, thus polarising society. Furthermore, the lack of reliable, neutral and accurate language when presenting news can trigger an increase in negative and violent users' reactions. It is necessary to find a linkage between disinformation and violent discourse in order to moderate online content, facilitate early detection of both phenomena and ensure healthy online behaviour. This research uses NLP techniques to explore the reliability of news headlines and their correlation with violent language generated in comments by users. In addition, the generation of a novel resource annotated in Spanish is created to jointly address the automatic detection of violent language and reliability.

Keywords

Natural Language Processing, Violent language, Reliability detection, Data analysis

1. Introduction

In our society, digitalisation exerts a profound influence with the Internet fundamentally reshaping how communication occurs and knowledge is obtained. This revolution has not only changed the way social connections are made but also transform the process of accessing new knowledge and staying informed in real-time. This information revolution has also boosted the risk of misuse of Information and Communication Technology (ICT) tools.

If negative interactions are already experienced in the physical world, they can spread even more easily in the virtual world due to the viral nature of content. This phenomenon has led to the proliferation of digital violence and the propagation of unreliable information.

It is crucial to comprehend the relationship between violence and disinformation in the digital environment. Disinformation and violence discourse have become essential areas of research. On one hand, combating the dissemination of fake news in the virtual realm is crucial to maintain information integrity and prevent the creation of societal alarms based on false facts. On the other hand, addressing the viral spread of violent content can mitigate significant consequences on a much broader population, since violent discourse, similar to disinformation, impacts reality perception and fosters a harmful atmosphere.

The notion of online violence extends beyond direct interactions and can manifest through information manipulation. The presence of digital violence in news can compromise their neutrality, resulting in the production of polarised and biased content that adversely impacts users. Such violence can contribute to the proliferation of unreliable information, casting doubt on reality and fostering mistrust in crucial areas such as public health or ideology. Research in these domains emerges as an essential tool to counteract these phenomena and foster a digital environment that is both safer and more trustworthy.

In [1] states that "in the civilised world, a language is a tool that enables the exchange of information and views, connects people of different nationalities, crosses religious and cultural barriers, allows knowledge, understanding, and helps build unity". However, if the language employed exhibits bias, manipulation, subjectivity, inaccuracy, or even aggression, the opposite effect may occur, potentially leading to societal fragmentation rather than cohesion.

Research has examined the detrimental impact of hate speech and disinformation on society, resulting in polarisation of society. By polarisation we mean "a social phenomenon characterised by the fragmentation of society into antagonistic factions with vehemently opposed values and identities that prevent cooperation and the search for a common good" [2]. Digital media have be-

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

beatriz.botella@dlsi.ua.es (B. Botella-Gil[1]); alba.bonet@dlsi.ua.es (A. Bonet-Jover[1]); rsepulveda@dlsi.ua.es (R. Sepúlveda-Torres[1]); patricio@dlsi.ua.es (P. Martínez-Barco[1]); stela@dlsi.ua.es (E. Saquete[1])

0009-0004-7094-6632 (B. Botella-Gil[1]); 0000-0002-7172-0094 (A. Bonet-Jover[1]); 0000-0002-2784-2748 (R. Sepúlveda-Torres[1]); 0000-0003-4972-6083 (P. Martínez-Barco[1]); 0000-0002-6001-5461 (E. Saquete[1])

come ideal venues for the propagation of violent and false content that generates group-based divisions in society.

As stated by [1], hate speech is based on the *divide et impera* concept, which intends to group society and turn different groups against one another. When society becomes polarised, disinformation can proliferate more easily. In particular, as reflected in [3]'s research, Spain stands out as the most polarised country in Europe.

The intersection of violent language and news disinformation in the digital world constitutes a crucial area of today's research. Addressing these issues will not only aid in identifying the reliability of information but also contribute to fostering a safer and more equitable online society. This research aims to conduct an exploratory study on the prevalence of violence in news comments, examining both the source and the subject matter. Additionally, it seeks to determine whether a correlation exists between the violent content generated by readers and the reliability of that content, achieved through an analysis of language usage. The objective is to determine whether the language used in a news article influences users to employ a higher level of violent language in the comments pertaining to the news item in question.

One of the limitations in both the task of detecting violent discourse and the task of detecting disinformation is the scarcity of resources in Spanish to train models in Natural Language Processing (NLP). For that reason, this research focuses on the analysis of a set of news items and comments in Spanish, in order to analyse violent and unreliable language in the Spanish language and propose a resource for both tasks.

The main novelty of our proposal consists of examining the correlation between violent discourse and disinformation in Spanish by means of NLP tools. To achieve this objective, the following contributions to this research area have been provided:

- A new resource consisting of headlines and news comments, annotated with the reliability and violence criteria. Reliability was manually annotated, while violence was automatically annotated by means of an existing automatic violence classifier originally created for tweets annotation that is applied in this research to news comments annotation.
- An exploratory analysis of how source profiles and topics can influence the generation of digital violence and exacerbate hostility among users.
- An assessment of how language used in news articles can impact the occurrence of hate speech, and more specifically, how the language used in news headlines may either incite readers to exhibit varying levels of violence when expressing their opinions on an issue.

The article is structured as follows: Section 2 shows the main research carried out in the fields of news reliability and language employed in disinformation, on the one hand, and violent language, on the other; in Section 3, the methodology employed for this exploratory study is outlined; Section 4 delves into the exploratory analysis to determine the relationship between disinformation and violence; Section 5 discusses the findings of this research, and finally, in Section 6, conclusions and limitations for future work are discussed.

2. State of the Art

The scientific community has approached the issue of online violence from various fields of knowledge, driven by the alarming prevalence of violent content in digital media. Efforts have been made to address issues such as hate speech, cyberbullying, and toxicity, aiming to early detect these problems and propose solutions.

Regarding the disinformation phenomenon, its objective is to disseminate misleading or unreliable content, which undermines societal trust by instilling insecurity and doubts regarding the authenticity of the information received.

This section aims to present relevant research related to both the violent discourse and the disinformation tasks, as well as to define the main important concepts of these lines of research, specially related to violent and reliable language.

2.1. Language in violence discourse

Language, as a means of communication, should be neutral. Still, its use can promote any ideology and incite hatred and violence [1].

An example of the power of language is evident in contexts of warfare, as seen in historical scenarios like the Cold War, the World War I, and the Vietnam and Iraq wars [4]. Similarly, in political discourses such as the 2016 US presidential elections or the Brexit referendum [5], language has played a significant role, particularly in addressing the issue of disinformation.

In the scientific field, research on linguistic violence has been approached from various angles. It is important to acknowledge the challenge in defining what constitutes violent content. Studies have aimed to delineate different forms of violent language prevalent in ICTs, including:

- **Ciberbullying**: the deliberate and repetitive use of specific ICTs like email, mobile phone messages, instant messaging, and personal online defamatory behaviour by an individual or group, with the intention of hostilely causing harm to another party [6].

- **Hate speech**: any form of communication that discriminates against an individual or group based on characteristics such as race, ethnicity, gender, sexual orientation, nationality, religion, or other identifying features [7].
- **Toxicity**: this term is defined by [8] and [9] as those messages that include unacceptable, rude, and disrespectful comments. They are messages of contempt that are part of what is called toxic discourse or toxic language, which either invite other users to leave the discussion or use language at the same level as the discussion.
- **Offensiveness**: in [10] asserts that this term is commonly defined as hurtful, derogatory, or obscene remarks directed from one person to another[1].

In addition, we also find other works that further specify the type of violence they study, such as in the case of misogyny or racism [11]. Our research will use the terms "violent language" or "violence" to refer to any type of message containing violent content.

Given the vast volume of data present in the virtual environment, manual detection of violent messages by humans is impractical. This underscores the significance of NLP tools in our research. From a NLP standpoint, detecting hate speech can be conceptualised as text classification. As stated by [12], "the automatic detection of this kind of speech is usually addressed as a classification task, and it is related to a family of other tasks such as detecting cyberbullying, offensive language, abusive language, toxic language, among others".

From this perspective, numerous research efforts focus on detection methods, including the development of resources aimed at aiding identification. These resources primarily consist of lexicons containing lists of negative words or expressions that may be present in messages, serving as indicators of potentially violent content. These word lists are employed in binary classification models to ascertain the presence or absence of hate speech [13], violent language [14], abusive language [15], and offensive language [16].

Furthermore, corpus generation is utilised in various violence detection strategies, thereby enhancing the model's capability to effectively discern both violent and non-violent content [17, 18].

Besides, efforts have been made to identify the most suitable detection tools that offer improved and expedited solutions to the problem. This includes exploring strategies such as heuristic techniques [19], Machine Learning (ML) [20], and Deep Learning (DL) [21].

[1] https://thelawdictionary.org/

2.2. Language in disinformation

Language also plays a very important role in disinformation detection and there are key linguistic indicators that are characteristic of news items. As described by [4], the essence of fake news or deceptive information is the news text, and, consequently, the language: "the news text is the basic communicative unit of journalism and consequently the basic unit of analysis in most research on fake news". First of all, it is important to define several key concepts in this field of research: fake news, disinformation, misinformation, reliability and veracity.

- **Fake news**: even if there is no universal definition, it is commonly defined as "any news that is suspected to be inaccurate, biased, misleading, or fabricated" and is understood as a "product of a range of practices that are related to the validity of information being shared by the news media" [4].
- **Disinformation vs. Misinformation**: in [1] states that "the information shared with malicious intent is recognised as disinformation, whereas the same information shared by a poorly informed party is considered as misinformation". The main difference lies in the intention: disinformation refers to that content which is intentionally and deliberately created with the intention to deceive while misinformation is inaccurate information that can results from an honest mistake, negligence or unconscious bias [22]. Briefly, as stated by [23], "disinformation is also wrong information but, unlike misinformation, it is a known falsehood".
- **Reliability vs. Veracity**: these concepts are closely related, but from what can be observed in the literature, the term veracity is usually used in tasks in which the information is contrasted and verified [24], while the concept of reliability is most commonly used in methods where the credibility of the source of the news is investigated, as is the case of the proposed source-based method [25].

For this research, we will focus on the disinformation and the reliability concepts. Reliability is an essential metric to be considered when assessing the quality of information. Some of the indicators that influence the reliability of a news item are: the ambiguity of the information, the lack of data and sources [26], the intention of hiding information, the representativeness or opacity of the headline, the external quotes from experts, the quotes from studies and organisations, emotional-charged expressions [27], or stylistic features such as the punctuation, the extension, the use of capital letters or informal or swear words [28].

In addition, [29] suggested that there are ten indicators that are most likely to detect the reliability of a message and these appear when the message is: complete, concise, coherent, well presented, objective and representative; when it contains no spin, uses expert sources, is perceived to have an impact and is professional.

As in Section 2.1, NLP plays an essential role in this line of research, as the continuous and easy access to the internet, the large volume of misleading content and its rapid viralisation make it impossible to process and treat data manually in the time required and before it becomes pervasive in society. For that reason, disinformation detection needs to be automated by means of NLP techniques.

Disinformation detection is specially being addressed as a classification task, by training supervised ML models to automatically distinguish between real and fake news [4]. For that purpose, annotated corpora are needed to mark patterns of language that will help in the training and classification processes. Most corpora generated in this domain is composed of news articles classified following several techniques, ranging from stylistic and linguistic annotations to binary classifications depending on fact-checking verification [28, 30, 26]. In addition, several methods are being applied to the disinformation problem based on knowledge, style, propagation, source or even hybrids methods, each of them approaching the automatic detection from different approaches [25].

2.3. Violence and Reliability: a new line of research

Polarisation emerges as a key factor linking the phenomena of disinformation (in particular, reliability) and violent online messages that initially appear disparate. Disinformation, by spreading erroneous or misleading information, can intensify polarisation by encouraging the generation of radical opinions and the adoption of hostile language. Polarisation, in turn, acts as a catalyst for the spread of hate speech, creating an environment conducive to distrust towards those with divergent views. This vicious cycle continuously feeds back on itself, with disinformation and hate speech fuelling polarisation, and with polarisation facilitating the spread of more disinformation and hate.

Through a detailed analysis of relevant case studies and research, it is assessed how this cycle contributes to the escalation of online violence and undermines the foundations of social cohesion and democracy [31]. It also discusses possible strategies to address these problems, highlighting the importance of media education, the promotion of critical thinking and the building of more inclusive and respectful online communities.

Among research linking hate speech and disinformation, it can be highlighted the work presenting by [12], which presents a dataset based on user responses to posts from Argentinian digital newspapers on Twitter. Their research is specially focused on the hate speech detection task but the authors work with replies to digital newspapers posts, which is an interesting point of view for our research because we also work with users' replies, in our case with news comments, but instead from digital newspapers posts, we focus on digital news.

These authors also describe in their article a work in progress on hate speech that analyses Spanish tweets linked to newspapers, which is also related to our work, albeit with a different type of document.

In [32] aim to analyse the relationship between the consumption of misinformation and the online hate and toxic language. To that end, they analyse a corpus of comments on Italian Youtube videos, distinguishing between four categories of hate speech and categorising two types of speech channels: questionable (channels likely to disseminate unverified and false content) and reliable (the remainder of the channels).

In [33] analyse a large Twitter dataset annotated with hate speech and counterhate speech. They state that, even if misinformation and hate speech have been tackled in parallel, they are often interwoven because "biased people justify and defend their hate speech using misinformation". That reason makes this linkage crucial for effective online content moderation.

To the best of the authors' knowledge, there is limited research that conducts an analysis connecting these two lines of research (disinformation and violent discourse), especially in the context of Spanish language. Furthermore, our research proposes a comprehensive analysis of both news reliability and violent discourse, using a newly created resource tailored specifically to address these two research areas.

3. Methodology

The main objective of this research is to ascertain the correlation between various attributes of news articles, including topic, source credibility, and content reliability, and the emergence of violence and polarisation within readers' comments on said articles. To achieve this goal, two principal analyses are conducted: i) an exploratory analysis into the prevalence of violent language within digital media, aiming to assess the degree of violence relative to topic and source; and ii) a study of the association between the reliability of news headlines and the occurrence of violent language within readers' comments on news articles. The methodology employed to accomplish this objective entails a dual process involving data collection and the analysis of both the violent language within comments and the reliability of news headlines. This is achieved through annotation using specific schemes

tailored for each aspect.

3.1. Data collection

News was gathered using a web crawler that extracted data from ten Spanish digital sources. A diverse sample of widely-read national media digital newspapers was selected, each representing various ideological orientations. This diverse selection enables us to assess the prevalence of violent language across different news sources. To analyse the extent of violence by topic, ten news items from each source were examined, resulting in a total of 100 articles for this initial study. These articles encompass a range of topics, ranging from politics, society, economics, health, and sports.

Given that news articles typically maintain a more neutral tone and considering that many of the chosen sources consist of newspapers authored by professional journalists, the decision was made to evaluate violence within the comments section rather than within the articles themselves. The objective is to gauge the level of violence stemming from readers' comments, as these comments reflect individual users' opinions and, for the most part, are not moderated, offering insights into how users express themselves.

Moreover, concerning disinformation, the focus of this study is placed on the content reliability, as unreliable content is deemed potential disinformation. In this regard, to investigate whether the reliability of the news items influences the level of present violence, news headlines were used to examine if there is a relationship between news reliability and violence in comments.

3.2. Violent language analysis

In this initial analysis, the primary objective is to examine the prevalence of violent language within the digital journalism landscape. To achieve this, news articles collected from various digital newspapers presented diverse political affiliations and levels of popularity. The selection criteria included not only the newspapers' popularity but also their editorial style and user engagement.

3.2.1. Data annotation

The annotation process of the downloaded comments was performed using an automatic violence classifier (anonymous). This classifier was developed by fine-tuning a RoBERTa model in Spanish, using a dataset of tweets annotated with Violent and Non-Violent labels. Given the similarity in the language and length between tweets and news comments, the system was applied to this research. Moreover, it yields significant results in discerning whether a tweet demonstrates violence, achieving an F_1m of 0.854 in a test set.

The chosen system classified a total of 1,757 comments as Violent and 3,940 comments as Non-Violent. Following the application of automatic classification, manual supervision was conducted by an expert in NLP specialised in criminology and violent language.

3.3. Reliability analysis

The objective of this second analysis is to identify a correlation between the usage of violent language in news comments and the reliability of news headlines. As detailed in (anonymous), the reliability of content is assessed by considering the objectivity and accuracy of the language used. Through this analysis, a research is conducted to determine whether a relationship exists between disinformation (unreliable news) and the hatred expressed in the comments pertaining to the news item in question.

3.3.1. Data annotation

The annotation task for this second analysis was conducted manually. Out of the 100 headlines, 31 were categorised as Unreliable, while 69 were categorised as Reliable. All annotations were performed by an expert in NLP specialised in linguistics and disinformation detection. The annotation of headlines considered the following reliability criteria, as outlined in (anonymous):

- **Data accuracy**: data should avoid vagueness or ambiguity. The employment of evasive or indefinite expressions suggests concealment or an inability to substantiate a fact. Additionally, the absence of evidence, such as scientific studies or verified official data, undermines the reliability of a news item.
- **Data objectivity**: the information presented should maintain neutrality. Information that sways the reader either positively or negatively, or that reflects the author's standpoint through personal remarks or experiences, indicates low reliability. Subjective data make the reader more vulnerable to believe unreliable news items.
- **Headlines style**: headlines should be informative, concise and neutral. Alarmist, subjective, opaque and striking headlines are characteristic of unreliable news items, as well as clickbait headlines, which are those "sensational headlines that often exaggerate facts, usually to entice readers to click on them" [34]. Other features, such as long headlines or the presence of many exclamation marks or words in capitals can also influence the reliability of the content.

4. Exploratory analysis

This section outlines three preliminary studies examining the use of violent language within the digital news landscape and its correlation with various external factors, including topic, source, and the reliability of the news content. These studies facilitate a comprehensive understanding of violent discourse behaviour within this context, as well as the impact of news language reliability on the proliferation of violent language in news comments.

4.1. Level and type of violence by topic

After analysing violent discourse related to the topics of the news items, the results obtained regarding the level of violence across different news topics are presented in Table 1.

Table 1
Violence percentage per news topic.

Topic	Violence
Politics	38.5%
Economics	28.6%
Society	27.0%
Sport	24.6%
Health	23.5%

As can be observed, political comments exhibit the highest level of violence (38.5%), followed by economics, society, sport and health, with the latter having the lowest percentage of violent language.

It is important to emphasise that while the news items align with these topics as categorised by digital media, comments that deviate from the associated topic of the news items have been encountered. For instance, economic news may also attract violence directed towards the political sphere due to the close relationship between these topics. Similarly, in sports news, sexist comments may arise, particularly when addressing women's football, as exemplified by:

- **Violent comment**: *Descanse en paz la z empoderada que quería hacer cosas de chicos, como jugar al fútbol y conducir su propio coche, con el riesgo que eso tiene. A trabajos y actividades de hombres, riesgos de hombres, eso es todo.* [Rest in peace to the empowered z who wanted to do boys' things, like play football and drive her own car, with all the risk that entails. Men's jobs and men's activities, men's risks, that's all.]

In addition, we found cases in different topics where the discussion drifts into radicalisation with comparisons to Nazi ideology, confirming what is known as Godwin's law, which states that as an online discussion lengthens, the probability of someone comparing another person or group of people to Nazis or Adolf Hitler tends to increase, regardless of the topic or position being debated [35]. For example:

- **Comment in political news item**: *socialistas? = autoritarios fascionazis!!* [socialists? = fascionazi authoritarians!!]
- **Comment in sport news item**: *Mucho machinazi llorando* [A lot of machinazi crying]

4.2. Level of violence by source

One of the factors we aimed to analyse was the percentage of violence generated by the source, contingent upon its credibility.

As stated by [25],"one can detect fake news by assessing the credibility of its source, where credibility is often defined in the sense of quality and believability". This research proposes to relate credibility to the content analysed according to the proposed reliability criteria in Section 3.3.1. Since there is currently no standard measure of media credibility, but several factors are taken into account, in this research we will consider as credible sources those presenting reliable content, while those sources whose news items analysed did not meet the reliability criteria of neutrality, objectivity, accuracy and coherence were classified as sources with low credibility. To accomplish this, a correlation will be established between the number of headlines annotated as reliable and unreliable, and the credibility level of the media in which these news items are published.

In this regard, three levels of credibility will be delineated: high credibility, indicated when the percentage of reliable headlines exceeds 80%; low credibility, identified when over 80% of the analysed headlines were deemed unreliable and failed to meet the established reliability criteria; and medium credibility otherwise, denoting sources that present both reliable and unreliable headlines in similar proportions. In our study, upon analysing all the news headlines, six sources were classified as highly credible, three were identified as having low credibility, and one source was deemed partially credible.

Results obtained concerning the level of violence according to the sources credibility are presented in Table 2.

From these results it can be concluded that the more credibility the source has, the lower the percentage of violence is generated in the comments associated with the news items published. Still, since this was a preliminary study to prove if there was a linkage between reliability and violence, the analysis was carried out in a small sample. As the results show that we can go even deeper

Table 2
Violence percentage per source credibility.

Sources credibility	Violent comments	Non-violent comments
High	28.42%	71.58%
Medium	34.57%	65.45%
Low	57.85%	42.15%

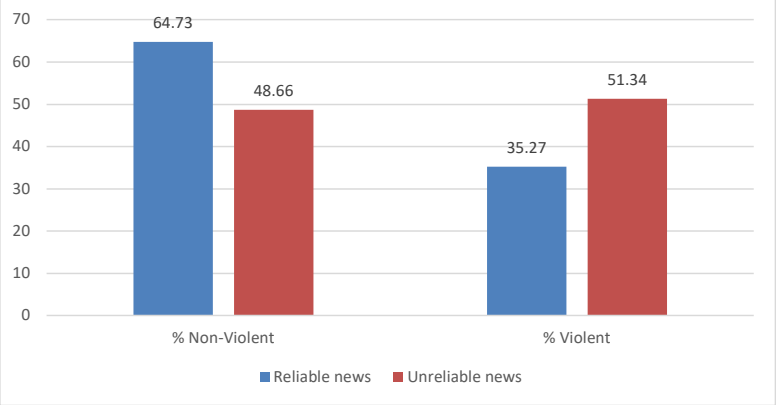

Figure 1: Degree of violence depending on the reliability of the headline.

to test our hypothesis, this analysis will be carry out in a larger sample in future work.

4.3. Degree of violence depending on the reliability of the headline

One of the main purposes of this research is to find if there is a relationship between the reliability concept (mostly related to disinformation tasks) and the violent discourse. This work aims to join these two lines of research to delve into the misleading and malicious digital content and propose a preliminary combined solution to automatic detect both disinformation and violent discourse. This analysis is carried out in the news scenario on the basis of the following data:

- **News**: 69 Reliable news (69%) and 31 Unreliable news (31%).
- **Comments**: 3,940 Non-Violent comments (69.15%) and 1,757 Violent comments (30.85%).

After conducting the analysis contrasting the violence generated in the comments of news articles with a reliable headline versus those with an unreliable headline, as depicted in Figure 1, it is evident that reliable headlines generate less violence than unreliable ones.

Following our analysis, we observed that of the 31 headlines classified as unreliable, 19 of them attracted a higher number of violent comments. This implies that more than 61.29% of the unreliable headlines elicited more negative responses from users, evidencing a relationship between the unreliability of a news item and the propensity of users to express violence in their comments.

On the other hand, of the 69 news items that have been classified as reliable, only 9 of these news items had more violent comments. Therefore, 81.15% of these news items recorded more non-violent comments and only 5.79% generated an equal amount of violent and non-violent comments. These findings support the reliability of the headlines, suggesting that those news items considered more reliable tend to generate a lower proportion of violent comments compared to the unreliable ones.

The following examples show how the way a headline is presented or written can incite users to generate more violent content in comments:

- **Unreliable headline**: *Pedro Sánchez en Estrasburgo; sentí vergüenza ajena* [Pedro Sánchez in Strasbourg; I felt ashamed]
  - **Violent comment**: *Usted y todos. Nos dejó a la altura del betún, como si fueramos una dictadura bolivariana cualquiera. (Que, sin duda, es lo que el quiere…)* [You and everyone else. He made us feel small, as if we

were a Bolivarian dictatorship (which, no doubt, is what he wants...)]

- **Unreliable headline**: *Detengamos a la izquierda ya* [Let's stop the left now]
  - **Violent comment**: *La izquierda no existe es todo extrema derecha los que se dicen de izquierdas en realidad no lo son, es una falsa izquierda que traicionando al pueblo se ha aliado con la banca usurera satanica que es la que hay que detener el verdadero enemigo de la humanidad* [The left does not exist it is all extreme right those who say they are leftists actually they are not, it is a false left that betraying the people has allied itself with the satanic usurious banking institution which is the real enemy of humanity that must be stopped]

As can be observed in the previous examples, the language employed in news headlines is inherently biased, characterised by a subjective style that mirrors the author's polarised political stance. The manner in which the headline is formulated and presented allows little space for readers to formulate their own conclusions, but rather prompts them to generate negative and violent content on the given topic.

5. Discussion

The present research has yielded a series of significant results, which are detailed below.

Topics: our analysis indicates that politics emerges as the subject with the highest incidence of violence in comments across various news topics. This finding suggests that the polarising and impassioned nature of politics encourages users to express themselves more freely, potentially leading to more confrontational and hostile interactions. Additionally, it is noteworthy that health garners the lowest frequency of violent comments, possibly due to its relatively lower prominence in our selection of news articles.

We have also found that there is not always a relationship between the topic of the news item and the type of violence that appears in users' comments. As explained in Section 4.1, we found cases of news items classified as economics or sports containing comments of sexist or political violence.

Credibility: the analysis also uncovers a direct association between the credibility of media sources and the prevalence of violence in online comments. Users are inclined to express a higher frequency of violent comments on news sourced from outlets that are perceived as less reliable. This observation suggests that the credibility of media sources plays a significant role in shaping the tone and content of user-generated comments. A lack of trust in the source can trigger negative and hostile reactions, underscoring the pivotal role of integrity and reliability in digital journalism.

Reliability: the reliability of the language used in a news item can indeed impact both the quantity and tone of the violent comments it provokes. Trustworthy media outlets typically uphold ethical and professional standards in information presentation, thus lowering the probability of users responding in such an aggressive manner. Conversely, media with a lower credibility level may disseminate misleading, biased, or sensationalist content, prompting negative and violent reactions from readers.

Additionally, it is crucial to highlight the significance of social media moderators used by certain media outlets. These moderators play a vital role in curbing the dissemination of violent comments by moderating and censoring those that contravene website usage policies.

6. Conclusions and future work

In this study, we have investigated the correlation between violence and disinformation. By integrating these two research approaches, our aim is to ascertain whether a relationship exists between violent language and the reliability language in news.

Our findings show that addressing both violence and reliability concurrently may facilitate the identification and mitigation of malicious digital content. For instance, the level of violence in comments may correlate with the reliability of news headlines and the credibility of the media outlets.

For future work, we aim to expand the initial resource created to propose a novel corpus annotated with reliability and violent language that facilitates the automatic detection of both tasks. The objective is to ensure a balanced distribution of the news articles across different topics in order to offer a more comprehensive and representative understanding of the diverse types of violence prevalent in the digital sphere.

In addition to expanding and balancing the corpus, a manual and meticulous revision process of the automatic violence annotation made by the classifier will be carry out, in order to ensure the correct classification of the comments. This revision task, which will be accomplished by two NLP experts in linguistics and violent discourse, will enable the generation of a quality annotated resource for both the violence discourse and disinformation tasks.

Finally, it is proposed to conduct an assessment of a set of comments categorised as Non-Violent to determine whether they present concealed violence, that is, whether

language employed subtly hides violence patterns, such in the case of irony or sarcasm. Alongside further examination of the extent of violence, a future hypothesis will focus on studying whether reliable news correlates with levels of mild and more subtler forms of violence, and whether unreliable news demonstrates more pronounced violence expressed in a more aggressive manner. This step will contribute to refining detection methods and enhancing the accuracy of evaluations.

In summary, this study represents a significant step towards a deeper understanding of the linkage between violence and disinformation in the NLP context, laying the groundwork for future research efforts. This research aims at fostering a safer and healthier digital environment by creating a resource that combines both news reliability and violent language of comments and thus proposing a common strategy to address these two research lines.

Acknowledgments

The research work is part of the R&D&I projects: CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; COOLANG.TRIVIAL: Technological Resources for Intelligent VIral AnaLysis (PID2021-122263OB-C22) funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; SOCIALFAIRNESS.SOCIALTRUST: Assessing trustworthiness in digital media (PDC2022-133146-C22) funded by MCIN/AEI/10.13039/501100011033/ and by the "European Union NextGenerationEU/PRTR". At regional level, this research has been funded by the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21) by the Generalitat Valenciana.

References

[1] M. Konieczny, Ignorance, disinformation, manipulation and hate speech as effective tools of political power, Policija i sigurnost 32 (2023) 123–134.

[2] A. J. Stewart, N. McCarty, J. J. Bryson, Polarization under rising inequality and economic decline, Science advances 6 (2020) eabd4201.

[3] N. Gidron, J. Adams, W. Horne, How ideology, economics and institutions shape affective polarization in democratic polities, in: Annual conference of the American political science association, 2018.

[4] J. Grieve, H. Woodfield, The language of fake news, Cambridge University Press, 2023.

[5] R. Greifeneder, M. E. Jaffe, E. J. Newman, N. Schwarz, The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation, Routledge, New York, NY, 2020.

[6] B. Belsey, Cyber-bullying: An emerging threat to the 'always on'generation, 2006, Retirado de: http://www. cyberbullying. ca/pdf/Cyberbullying_Article_by_Bill_Belsey. pdf (2014).

[7] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, 2012, pp. 19–26.

[8] R. Nielsen, N. de Domenico, Volume and patterns of toxicity in social media conversations during the covid-19 pandemic (2020).

[9] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 1391–1399.

[10] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of abusive language: the problem of biased datasets, in: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 602–608.

[11] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–19.

[12] J. M. Pérez, F. M. Luque, D. Zayat, M. Kondratzky, A. Moro, P. S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, et al., Assessing the impact of contextual information in hate speech detection, IEEE Access 11 (2023) 30575–30590.

[13] G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 1980–1984.

[14] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & internet 7 (2015) 223–242.

[15] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[16] F. M. Plaza-del Arco, A. B. P. Portillo, P. L. Úda, B. Gil, M.-T. Martín-Valdivia, SHARE: A lexicon of harmful expressions by Spanish speakers, in:

Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1307–1316.
[17] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata, A multilingual evaluation for online hate speech detection, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–22.
[18] V. Kolhatkar, H. Wu, L. Cavasso, E. Francis, K. Shukla, M. Taboada, The SFU opinion and comments corpus: A corpus for the analysis of online news comments, Corpus Pragmatics 4 (2020) 155–190.
[19] F. Huang, H. Kwak, J. An, Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 90–93.
[20] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, P. Nakov, A large-scale semi-supervised dataset for offensive language identification, arXiv preprint arXiv:2004.14454 (2020).
[21] C. Arcila-Calderón, J. J. Amores, P. Sánchez-Holgado, D. Blanco-Herrero, Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in Spanish, Multimodal technologies and interaction 5 (2021) 63.
[22] D. Fallis, The varieties of disinformation, The philosophy of information quality (2014) 135–161.
[23] B. C. Stahl, On the difference or equality of information, misinformation, and disinformation: A critical research perspective, Informing Science 9 (2006) 83.
[24] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.
[25] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (2020) 1–40.
[26] S. Mottola, Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español, Discurso & Sociedad (2020) 683–706.
[27] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al., A structured response to misinformation: Defining and annotating credibility indicators in news articles, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 603–612.
[28] B. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 759–766.
[29] A. Appelman, S. S. Sundar, Measuring message credibility: Construction and validation of an exclusive scale, Journalism & Mass Communication Quarterly 93 (2016) 59–79.
[30] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.
[31] G. Americans, Medición del impacto de la información falsa, la desinformación y la propaganda en américa latina, Global Americans (2021).
[32] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. K. Novak, F. Zollo, Dynamics of online hate and misinformation, Scientific reports 11 (2021) 22083.
[33] J. Y. Kim, A. Kesari, Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic, Journal of Online Trust and Safety 1 (2021).
[34] S. Chawda, A. Patil, A. Singh, A. Save, A novel approach for clickbait detection, in: 2019 3rd International conference on trends in electronics and informatics (ICOEI), IEEE, 2019, pp. 1318–1321.
[35] F. A. Wilson, Enough Already!: A Socialist Feminist Response to the Re-emergence of Right Wing Populism and Fascism in Media, Brill, 2020.