Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09678-9

ORIGINAL PAPER

RUN-AS: a novel approach to annotate news reliability for disinformation detection

Alba Bonet-Jover[1] · Robiert Sepúlveda-Torres[1] · Estela Saquete[1] · Patricio Martínez-Barco[1] · Mario Nieto-Pérez[1]

Accepted: 13 June 2023

Abstract
The development of the internet and digital technologies has inadvertently facilitated the huge disinformation problem that faces society nowadays. This phenomenon impacts ideologies, politics and public health. The 2016 US presidential elections, the Brexit referendum, the COVID-19 pandemic and the Russia-Ukraine war have been ideal scenarios for the spreading of fake news and hoaxes, due to the massive dissemination of information. Assuming that fake news mixes reliable and unreliable information, we propose RUN-AS (Reliable and Unreliable Annotation Scheme), a fine-grained annotation scheme that enables the labelling of the structural parts and essential content elements of a news item and their classification into Reliable and Unreliable. This annotation proposal aims to detect disinformation patterns in text and to classify the global reliability of news. To this end, a dataset in Spanish was built and manually annotated with RUN-AS and several experiments using this dataset were conducted to validate the annotation scheme by using Machine Learning (ML) and Deep Learning (DL) algorithms. The experiments evidence the validity of the annotation scheme proposed, obtaining the best $\mathbf{F_1 m}$, 0.948, with the Decision Tree algorithm.

Keywords Natural language processing · Annotation guideline · Dataset annotation · Reliability detection · Disinformation detection

1 Introduction

In turbulent times, disinformation becomes a great enemy. Shu et al. (2020) define disinformation as fake or inaccurate information that is intentionally spread to mislead and/or deceive. When it comes to political, social and health issues, factors such as disorder, fear and economic or ideological interests increase the volume of

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, Patricio Martínez-Barco and Mario Nieto-Pérez have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 06 August 2023

disinformation. This global social problem is part of our lives to such an extent that specific terms are created to refer to it, as in the case of “infodemic”, used by the World Health Organization to refer to the excess of false or misleading information during a disease outbreak, as occurred during the COVID-19 pandemic, or “post-truth”, defined by the Cambridge Dictionary[1] as those situations in which people are more likely to accept an argument based on their emotions and beliefs, rather than one based on facts.

The Internet has fuelled the need to be continuously informed and that thirst for information results in a faster dissemination of unverified news, as anyone can share and access information at no cost. For that reason, the disinformation phenomenon has become a challenge for many researchers from different research areas. In Natural Language Processing (NLP), several approaches are used to tackle this problem, such as automated fact-checking, sentiment analysis, deception and stance detection, contradiction detection, credibility, among others. Even if these lines of research address the problem from different perspectives, they are complementary, since they often share common objectives and problems (Saquete et al., 2020).

As stated by Giansiracusa (2021), “whether we want it or not, automation is coming to journalism, and none are more poised to take advantage of this than the peddlers of fake news”. A complex mix of cognitive, social and algorithmic biases makes us more vulnerable to believing online disinformation and being manipulated (Shao et al., 2017). For that reason, it is important to combat disinformation in the same environment in which it is generated: the digital world. The huge amount of disinformation and its rapid dissemination has led computer scientists to automate tasks and develop computational models, since it is impossible to process and manually analyse such a large amount of data within the necessary short time frames. Expert intervention and the use of algorithms are potential solutions to tackle the disinformation problem. Human intervention is essential to provide the expertise and the examples to train, as well as to check the information and supervise the decisions made by the system trained, while assisted systems are needed to automate tasks that experts would not be able to carry out manually. Algorithms are responsible for the spreading of disinformation, but also for its mitigation, so the cause of this phenomenon may in turn be the solution to the problem. However, these algorithms are not yet robust enough to perform a verification of which information is false or true (Figueira and Oliveira, 2017).

To address the disinformation problem, computational systems need labelled examples to train, since “manually-curated gold standard annotations are a prerequisite for the evaluation and training of state-of-the-art tools for most Natural Language Processing (NLP) tasks” (Stenetorp et al., 2012). The problem lies in the fact that annotating corpora is a costly, slow and time-consuming task, so labelled corpora are scarce, especially in languages other than English, such as Spanish.

As mentioned above, several lines of research are working to solve the disinformation problem, commonly known as fake news detection. In this research, instead of focusing on the veracity concept (fake news detection), we deal with the concept

[1] https://dictionary.cambridge.org/es/.

of reliability (reliable and unreliable news detection). Both concepts are closely related, as fake news is news created for a specific purpose that includes both reliable and unreliable information but, since we are tackling the problem from a linguistic and semantic perspective, an absolute judgment on the veracity of a text is not possible without the use of world knowledge to corroborate the information. However, it is possible to detect when a text has the appearance of being unreliable, and therefore to generate doubt in the reader's mind about its veracity. Before presenting our proposal, it is important to differentiate these two concepts. In this research, the term reliability refers to the quality of being credible. As defined in the Oxford dictionary,[2] reliability is "the quality of being likely to be correct or true". Another similar entry is provided by the Collins dictionary[3] that states that "information that is reliable or that is from a reliable source is very likely to be correct". Concerning the term veracity, both dictionaries define it as "the quality of being true". Taking these definitions into account, it can be noted that the difference between both terms lies in the word "likely" included in the definition of reliability.

For detecting the veracity of news, information verification (such as fact-checking) is needed to decide if a given statement is true of false. This classification depends on several factors, including external knowledge to contrast the information, therefore it cannot be determined by only taking into account language, textual characteristics or semantics. However, this work is not focused on whether a news item is true or false, but rather to determine if there is sufficient evidence to consider the news item credible or not. By virtue of this, we focus on the stipulated difference between veracity and reliability.

The novelty of our proposal is the design of an accurate and innovative semantic annotation scheme that focuses on classifying news as Reliable or Unreliable from a linguistic perspective and without external knowledge. This proposal incorporates two well-known journalistic techniques: (i) the Inverted Pyramid, focused on structuring a news item in a clear way and on providing all the information in order of relevance and (ii) the 5W1H, a technique allowing to present the content in a complete and precise manner by answering six key questions (WHO, WHAT, WHEN, WHERE, WHY and HOW).

The annotation proposal, hereafter referred to as RUN-AS (Reliable and Unreliable News Annotation Scheme), enables the essential parts of a news item to be detected, namely the structure (Inverted Pyramid) and the content (5W1H) along with the reliability of its semantic elements. To annotate the reliability of the labels, the accuracy and neutrality of the information provided is taken into account, as well as the presence of personal remarks, derogatory language, emotionally charged expressions, lack of scientific evidence or language that has a negative or positive influence on the news and which usually has a specific intention, such as persuasive or exhortative expressions (all these characteristics are explained in Sect. 3.3). The strength of this work is a novel annotation guideline that can be used to detect reliable and unreliable information for its future application in disinformation detection

[2] https://www.oxfordlearnersdictionaries.com/.

[3] https://www.collinsdictionary.com/.

tasks. The reliability detection may enable an overall approach towards the news item by only reading the text and before taking the time to verify the information in reliable sources. This could be useful as an initial step for detecting suspicious information, thereby supporting not only users but also journalists by helping them to quickly detect disinformation.

This paper is structured as follows: Sect. 2 presents an overview of the most relevant scientific literature; Sect. 3 describes the annotation scheme proposed and the several criteria followed to classify the reliability of news; Sect. 4 introduces the dataset created to test our proposal and two inter-annotator agreements to avoid bias in assessing news; Sect. 5 presents several experiments that validate our annotation scheme; Sect. 6 summarises the results and discussion; Sect. 7 presents an experiment to automatically detect the reliability of the 5W1H elements; and finally Sect. 8 presents the conclusions of this research and future work.

2 Related work

This section presents some relevant literature that helped us to delimit our research and to analyse recent annotation schemes and datasets related to disinformation. In this field, most datasets focus on annotating the whole news as true or false, as shown in subsection 2.1. Furthermore, since our proposal aims not only to annotate the whole news but also the individual parts and semantic elements based on journalistic techniques, some state-of-the-art work that focuses on those techniques is presented in subsection 2.2. Finally, given that our proposal centres on analysing linguistic characteristics of news in order to detect their reliability, section 2.3 describes some research that studies these characteristics.

2.1 Annotated corpora for disinformation detection

Several datasets have been released to train computational models created for disinformation detection.

Two noteworthy corpora that focus on deception detection are the LIAR dataset, comprising 12,836 real-world short statements (Wang, 2017), and the EMERGENT dataset (Ferreira and Vlachos, 2016) containing 300 claims and 2,595 associated news articles. Regarding the annotation of these corpora, the LIAR dataset presents a scale of six fine-grained labels (pants-fire, false, barely-true, half-true, mostly-true and true), while the EMERGENT corpus classifies news into three veracity values (true, false and unverified) and assigns a stance label to the headline with respect to the claim (for, against and observing). Vlachos and Riedel (2014) also released a fake news detection and fact-checking dataset comprising 221 statements collected from PolitiFact[4] and Channel 4[5] and annotated with a five-label-tag classification:

[4] http://www.politifact.com/.

[5] http://blogs.channel4.com/factcheck/.

True, MostlyTrue, HalfTrue, MostlyFalse and False. Even if these datasets have been created specially for the context of fact-checking, they are also useful for deception detection and include news annotated to train.

We can also highlight the CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection (Shahi et al., 2021), a lab that focuses on evaluating technology that enables, especially in this task and particularly in subtask 3A, automatic detection of the news story's veracity. Given the text and the title of a news article, the goal was to predict whether the main claim of the article was true, partially true, false, or other. The participants were provided with a training dataset consisting of 900 news articles, leaving 354 articles for testing.

Another study focused on automatically identifying fake content in online news (Pérez-Rosas et al., 2017). These authors introduce two new datasets for fake news detection covering several domains and linguistic differences between legitimate and fake news articles were analysed. Regarding the construction of the datasets, the first one was obtained through crowdsourcing (240 legitimate news and 240 fake news) while the second dataset was collected from web sources (100 legitimate news and 100 fake news). To evaluate human ability in detecting fake news and the accuracy of their system, they created an annotation interface and asked annotators to label the developed datasets by choosing between "Fake" or "Legitimate" according to their perceptions. Their system performed well, even outperforming humans.

The pandemic has produced much disinformation, which has triggered new lines of research and datasets focused on COVID-19. As our dataset is also focused on health and COVID-19, it is relevant to mention two recent corpora addressing this domain: a fake news dataset consisting of 10,700 fake and real news annotated into real or fake (Patwa et al., 2021) and a large COVID-19 Twitter Fake News dataset (CTF), introduced by Paka et al. (2021), which works with labelled and unlabelled tweets using two-scale labels (fake and genuine).

Concerning corpora in other languages, Spanish resources are scarce, creating a need for proposals that focus on the Spanish language. A dataset created for studying automatic fake news detection in Spanish was released by Posadas-Durán et al. (2019), consisting of 491 true news and 480 fake news annotated with two labels (real and fake). Regarding disinformation corpora in languages other than English or Spanish, a dataset of labeled true and fake news in Portuguese called the Fake. Br corpus was presented by Silva et al. (2020). It is composed of 7200 news (3600 fake and 3600 legitimate news). To construct the corpus, for each fake news, a corresponding true news was collected, topically related, thus obtaining a corpus of aligned true and fake news.

Assaf and Saheb (2021) present a novel dataset of Arabic fake news containing 323 articles (100 reliable news and 223 unreliable news) and focused on traditional linguistic features. The news was manually collected by journalists and annotated by two human experts, whose agreement was measured through Cohen's Kappa. Furthermore, to study the differences between news articles from reliable and unreliable sources, Gruppi et al. (2018) constructed two datasets of political news articles from United States sources (997 reliable, 794 unreliable and 50 satire) and Brazilian sources (4698 reliable, 755 unreliable and 58 satire). For each article, they computed every feature on title and body text separately from a set of significant features in both languages and assigned a class Reliable (R), Unreliable (U) or Satire (S) based on the source from which the article was collected. Then, they constructed a set of roughly equivalent sets of features in both languages and classified them into 4 categories (complexity, style, linguistic and psychological). This study allowed them to show that differences exist between news articles from reliable and unreliable sources. What differentiates these previously mentioned two corpora from the rest is their classification into reliable and unreliable, an annotation that is closer to our approach and which is hardly used.

To the authors' knowledge, most current datasets classify and annotate news with a single global veracity value (true or fake, even if some of them propose a scale of veracity degrees or stance degrees). Many datasets created for disinformation detection have so far focused on fact-checking techniques, veracity classification (true/false) and global news annotation.

2.2 Corpora based on the journalistic techniques

Considering that our proposal uses journalistic concepts such as the Inverted Pyramid and the 5W1H, this subsection focuses on presenting some corpora that also use them. Many of these studies are using event extraction or semantic role labeling tasks. The following datasets can contribute to the background of our research because they focus on the journalistic techniques that are the basis of this research. Norambuena et al. (2020) propose the Inverted Pyramid Scoring method to evaluate how well a news article follows the inverted pyramid structure using main event descriptors (5W1H) extraction and news summarisation. Their proposal, which was evaluated in a dataset consisting of 65,535 articles from the Associated Press News (AP News), shows that the method adopted helps to distinguish structural differences between breaking and non-breaking news, reaching the conclusion that breaking news articles are more likely to follow the inverted pyramid structure. Another interesting work related to the 5W1H journalistic concept is that of Chakma and Das (2018), in which an annotation approach to assign semantic roles is described. This proposal is not applied to news, but to a corpus of tweets related to the US elections of 2016. To annotate the 5W1H, a Question and Answer (QA) approach was used to extract the answers to those questions and a corpus of 3000 tweets randomly sampled was used for this research. Furthermore, Khodra (2015) introduces a new 5W1H corpus of Indonesian news articles to train event extraction. The corpus, consisting of 90 news items obtained from popular news websites, was manually labelled by three human annotators following the 5W1H concept and extracting the event information of the news item.

These journalistic techniques are used in our dataset to detect structural parts, semantic events and linguistic patterns that can help to classify the reliability of news. To describe semantic events, all the elements related to the 5W1H questions are annotated. The 5W1H is usually used to detect the main event of a story that is usually found at the beginning of the story, in the title or lead of a news item. Our proposal aims to not only annotate the main 5W1H of the story appearing in the opening paragraphs, but all the 5W1H located in the article and related to other events or ideas. Disinformation can be found in any part of a news item and in any sentence or idea of the story, not only in the main event. The novelty of our annotation compared to the state of the art lies in the annotation of the 5W1H of all the parts of a news item (from the title to the conclusion), permitting more in-depth analysis of the whole news article.

2.3 Research focused on linguistic characteristics to detect disinformation

In this subsection we want to highlight relevant research focused on analysing linguistic characteristics in news to determine credibility. This point is relevant to our research because our annotation aims to mark disinformation patterns through language and textual characteristics.

Zhang et al. (2018) present a set of content and context indicators for article credibility. Regarding the content indicators, which are the ones that are of interest to our research, this work introduces some indicators that can be determined by analysing the title and text of the article without consulting outside sources or metadata. These indicators are: title representativeness, clickbait title, quotes from outside experts, citation of organizations and studies, calibration of confidence, logical fallacies, tone and inference. The authors introduce a dataset of 40 articles annotated with both content and context indicators. Furthermore, Horne and Adali (2017) state that the style and the language of articles allows differentiation of fake from real. This study is conducted in three separate datasets (containing real, fake and satire news) and analysed via three content based features categories: stylistic, complexity, and psychological. By studying similarities between news, they show that there is a notable difference in titles and content between fake and real news in terms of length, punctuation, quotations, lexical features or capitalized words. For stylistic and psychological features, the authors used the Linguistic Inquiry and Word Count (LIWC) dictionaries (Pennebaker et al., 2015), which is a text analysis program that counts words in psychologically meaningful categories and is available in different languages.

Another study showing that linguistic characteristics can help determine the truthfulness of text is that of Rashkin et al. (2017). This work compares the language of real news with that of satire, hoaxes and propaganda. To analyse the linguistic patterns, they sampled standard trusted news articles from the English Gigaword corpus and crawled articles from seven different unreliable news sites. Mottola (2020) also carries out a comparative study between Italian and Spanish in order to identify the common textual characteristics of digital disinformation. To that end, the author introduces a corpus made up of fake news published on digital platforms by both Italian and Spanish users and recognised as fake by two well-known fact-checking agencies: Bufale un Tanto Al Chilo[6] and Maldita.[7] Through this linguistic analysis, it is shown that there are several

[6] https://www.butac.it/

[7] https://maldita.es/.

characteristics that fake news share related to headlines, punctuation, capital letters, lack of data or emotional aspects.

We have referred to these linguistic studies because our proposal shares many of the same linguistic features, which are detailed in Sect. 3.

Our proposal makes a threefold contribution to disinformation detection. Firstly, as explained above, this research proposes a reliability classification instead of a veracity rating which is a novel way of classifying news that considers its textual and linguistic features, without reverting to external knowledge. Secondly, state-of-the-art corpora annotate news with a single global veracity value, whereas our proposal aims to annotate all the structural parts and semantic elements of a news item more precisely, by taking into account linguistic/textual characteristics and following journalistic techniques. Finally, this fine-grained annotation produces a quality resource in Spanish, which is essential to train and make progress in NLP. Given the premise that fake news is unreliable news that usually mixes true and false information, determining the reliability of the essential parts separately may help to determine the global reliability of the news item. Moreover, identifying which parts or elements have a greater influence in determining the reliability of a news item will serve to justify the final decision.

3 RUN-AS: annotation scheme based on journalistic techniques

3.1 Annotation labels

The goal of this annotation proposal is to analyse news on the basis of a purely textual and linguistic analysis to find out whether the way in which a news item is structured or written influences its reliability. The classification into Reliable or Unreliable may help to generate a report justifying that decision so that, at a later stage, it can be verified with fact-checking techniques. News has been annotated with Brat, an intuitive web-based annotation tool that integrates NLP technology (Stenetorp et al., 2012). RUN-AS (Reliable and Unreliable News Annotation Scheme) is based on two well-known journalistic techniques: the Inverted Pyramid and the 5W1H.

To find out whether a news item presents objective information and follows journalistic standards, this proposal enables a three-level annotation: Structure (Inverted Pyramid), Content (5W1H) and Elements of Interest. Regarding the structure, each label is described in detail and no examples are needed to clarify them, as the Inverted Pyramid is an intuitive annotation and it is easy to understand which part each label refers to. However, concerning the 5W1H and the Elements of Interest, the semantics of these levels makes the annotation more complex and subjective, so, in addition to the description, examples in Spanish of each label taken from the dataset are provided below along with the English translation in italics and in square brackets.

3.1.1 Structure labels

The Inverted Pyramid structure is one of the techniques used by journalists to reflect objectivity in a news item (Thomson et al., 2008). It consists of presenting the information in order of relevance, placing the most relevant information at the beginning and the least important at the end. This structure allows "users to quickly acquire key story points" and to "better facilitate information processing" (DeAngelo and Yegiyan, 2019). The five structure labels of our proposal are TITLE, SUBTITLE, LEAD, BODY and CONCLUSION. The annotation of these parts provides information on whether or not the news item follows the standard of the journalistic structure. In terms of structure, it is assumed that each source has its own particular style of writing and that not all parts have to be present (such as the SUBTITLE or the CONCLUSION). However, the lack of essential parts of a news item (such as the TITLE, the LEAD or the BODY) strongly suggests that a news item is poorly structured. At structure level, the main parts of the Inverted Pyramid hypothesis are annotated:

- TITLE: sentence containing the main idea of the news article and summarising the essential information of a story.
- SUBTITLE: sentence completing or repeating the information of the TITLE or providing additional information.
- LEAD: main and first paragraph presenting the essential information of the news article by answering the six key questions of the 5W1H concept. It develops and usually repeats the idea presented in the headline (Thomson et al., 2008).
- BODY: set of paragraphs developing the story and presenting in detail all the main and additional information regarding the 5W1H.
- CONCLUSION: last sentence or paragraph summarising the content of the news article. It does not present new information and is not always present.

3.1.2 Content labels

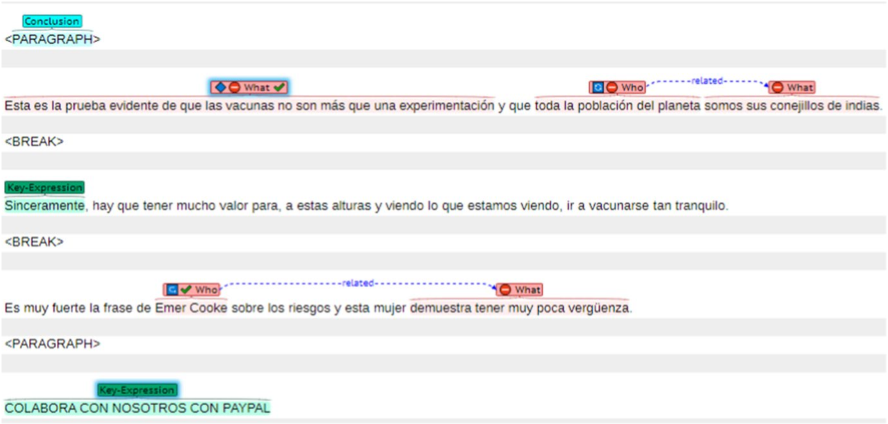
The other technique used by journalists to write a news item accurately and completely is the 5W1H which consists of answering six key questions. These questions describe the main event of a news story (Hamborg et al., 2018) and are usually found at the beginning of the news item, such as the TITLE or the LEAD. As stated by Chakma et al. (2020), "the 5W1H represents the semantic constituents of a sentence which are comparatively simpler to understand and identify". If a news item answers all these questions, it will mean that the information is communicated in a complete way and, therefore, the news item will have a higher degree of credibility than a news item that does not communicate the information in such a precise way. At content level, the annotation marks the events of the news according to the 5W1H. As defined by Hordofa (2020), an event "is a natural way to explain complicated relations between people, places, actions and objects" but it is also the natural way to describe the news, the way in which consumers understand what happened in the world (Hou et al., 2015). To describe this event, all the semantic elements related to the 5W1H questions are annotated as Reliable or Unreliable depending on their level of accuracy and objectivity.

- WHAT: facts, circumstances, actions.
  < WHAT > Los contagios por coronavirus se disparan< /WHAT >
  [*coronavirus infections skyrocket*]
- WHO: subject, entity.
  < WHO > La Agencia Europea del Medicamento< /WHO >
  [*European Medicines Agency*]
- WHEN: time, moment.
  < WHEN > El 21 de diciembre< /WHEN >
  [*On 21 December*]
- WHERE: place, location.
  < WHERE > En España< /WHERE >
  [*In Spain*]
- WHY: cause, reason
  < WHY > A causa de la enfermedad< /WHY >
  [*Due to the disease*]
- HOW: manner, method
  < HOW > Con abundantes vaporizaciones< /HOW >
  [*With abundant vaporisations*]

3.1.3 Elements of interest labels

The level of Elements of Interest enables the marking of textual information that could distinguish unreliable from reliable news:

- KEY_EXPRESSION: phraseology that urges readers to share the information or that expresses emotions such as fear, contempt, alarm or economic purposes.
  < KEY_EXPRESSION > Vamos a salvar vidas compartiendo esta gran información< /KEY_EXPRESSION >
  [*Let’s save lives by sharing this important information*]
- FIGURE: feature that can be verified by fact-checking techniques.
  < FIGURE > 15< /FIGURE > pacientes han dado positivo
  [*15 patients tested positive*]
- QUOTE: label that marks the presence of quotes in the news item.
  El experto cree que es solo < QUOTE >“cuestión de tiempo”< /QUOTE >
  [*The expert thinks it is only “a matter of time”*]
- ORTHOTYPOGRAPHY: when writing is poor and the text contains grammatical, spelling or formatting mistakes.
  < ORTHOTYPOGRAPHY >¿‘Porké?< /ORTHOTYPOGRAPHY >
  [*Whi*?]



<PARAGRAPH>

Esta es la prueba evidente de que las vacunas no son más que una experimentación y que toda la población del planeta somos sus conejillos de indias.

<BREAK>

Sinceramente, hay que tener mucho valor para, a estas alturas y viendo lo que estamos viendo, ir a vacunarse tan tranquilo.

<BREAK>

Es muy fuerte la frase de Emer Cooke sobre los riesgos y esta mujer demuestra tener muy poca vergüenza.

<PARAGRAPH>

COLABORA CON NOSOTROS CON PAYPAL

Fig. 1 Annotation of 5W1H, Inverted Pyramid and Elements of Interest on Brat

An example of the annotation in Brat can be observed in Fig. 1[8] and the translation into English of the text is presented below: *This is clear proof that vaccines are nothing more than experimentation and that the entire population of the planet is being used as guinea pigs. Sincerely, one has to have a lot of courage, at this point, seeing what we are seeing, to calmly go and get vaccinated. Emer Cooke's statement about the risks is very strong and this woman shows very little shame. COLLABORATE WITH US VIA PAYPAL.*

3.2 Attributes

Besides annotating Structure, Content and Elements of Interest labels, our annotation scheme includes several attributes for these labels, with specific values that help to provide essential information in the annotation:

reliability is the main attribute of our annotation and allows classification of each element as well as the global news item with the values Reliable or Unreliable, depending on the level of accuracy, objectivity and the linguistic characteristics (the reliability criteria are explained in detail in Sect. 3.3).

main_event is only used with the WHAT label and indicates the main event(s) of the story. The event describes all the semantic elements involved in the story, that is, all the 5W1H. It is possible to find several events (each one with its own 5W1H), since news communicate not only the main idea, but also secondary events. Our objective is not only to annotate the main 5W1H (i.e. only 6 elements in the whole news item), but as many 5W1H as possible that are of interest in the story. This attribute has no values.

[8] https://www.eldiestro.es/2021/04/grandisimo-escandalo-la-ema-ve-vinculos-entre-astrazeneca-y-los-coagulos-y-trombosis-pero-dice-que-el-riesgo-de-mortalidad-por-covid-es-mucho-mayor-que-el-riesgo-de-mortalidad-por-estos-efect/.

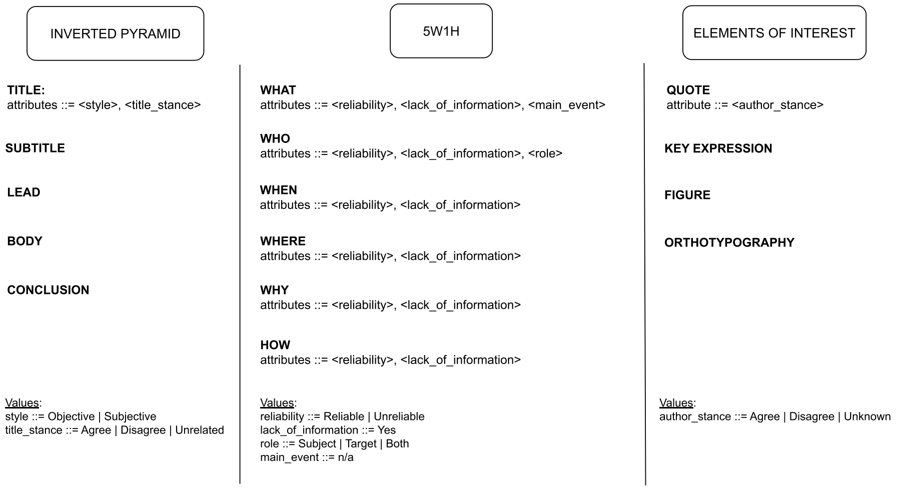

INVERTED PYRAMID
TITLE:
attributes ::= <style>, <title_stance>
SUBTITLE
LEAD
BODY
CONCLUSION
Values:
style ::= Objective | Subjective
title_stance ::= Agree | Disagree | Unrelated
5W1H
WHAT
attributes ::= <reliability>, <lack_of_information>, <main_event>
WHO
attributes ::= <reliability>, <lack_of_information>, <role>
WHEN
attributes ::= <reliability>, <lack_of_information>
WHERE
attributes ::= <reliability>, <lack_of_information>
WHY
attributes ::= <reliability>, <lack_of_information>
HOW
attributes ::= <reliability>, <lack_of_information>
Values:
reliability ::= Reliable | Unreliable
lack_of_information ::= Yes
role ::= Subject | Target | Both
main_event ::= n/a
ELEMENTS OF INTEREST
QUOTE
attribute ::= <author_stance>
KEY EXPRESSION
FIGURE
ORTHOTYPOGRAPHY
Values:
author_stance ::= Agree | Disagree | Unknown

Fig. 2 RUN-AS Guideline

role is the attribute used with the WHO label. It indicates the role played by the subject/entity of the event. The role attribute presents 3 values: Subject (if the entity causes the event), Target (if the entity receives the effects of the event) and Both (if the entity performs both functions).

lack_of_information is used with the 5W1H labels to indicate if scientific evidence or important data is missing. This attribute has a single value (Yes), indicating when such evidence is missing.

title_stance is only used in the structure annotation of the TITLE. It serves to indicate the relation and level of consistency between the TITLE and the BODY of a news item by means of the following values: Agree (information is consistent), Disagree (information is inconsistent) or Unrelated (information has no relation).

style is an attribute, which, as with the title_stance is only used in the TITLE, but in this case marks the values Objective or Subjective of the information provided in the TITLE.

author_stance is used with the QUOTE label and it serves to annotate the author's stance, represented by the following values: Disagree (to express its disagreement towards the idea), Agree (to share its agreement) or Unknown (just to inform, without showing its stance towards it).

Figure 2 shows a summary of the different labels of the annotation scheme with the specific attributes of each label, and the possible values for each attribute.

3.3 Reliability criteria

This work focuses on assigning a reliability value to the essential content labels described in our annotation scheme. The complexity of the task is detecting patterns of disinformation without corroborating information against external sources, and only taking into account textual and linguistic characteristics, as well as style. This makes the annotation more subjective because the analysis depends on factors such as the author's writing style and purpose, the emotional charge present in the language, or how gullible the reader is when reading news.

Despite this subjectivity, there are textual and linguistic features that enable detection of the reliability of a news item and an evaluation of all the Reliable and Unreliable elements of each part of the news item, permitting an assessment of the news item's overall reliability. The criteria taken into account when classifying the reliability of semantic elements are: accuracy of the content elements (5W1H), objectivity (reflected in the Elements of Interest features), titles, personal remarks and lack of information. To explain the criteria established by our scheme when classifying reliability, we will rely on some research mentioned in Sect. 2.3, whose characteristics are similar to the ones used in this work.

- **Accuracy**: for the information provided to be Reliable, it is important that the data is accurate and does not leave room for vagueness or ambiguity. Evasive or vague expressions indicate that something is being concealed or that a fact cannot be justified, which makes the information provided Unreliable. For example, it is more reliable to give an exact date or precise details on a scientist (name, institution, degree) than to generalise or to provide inaccurate data.

  < WHENreliability: := Unreliable > Hace mucho tiempo< /WHEN >
  [*A long time ago*]
  < WHENreliability: := Reliable > El viernes 19 de marzo< /WHEN >
  [*On Friday 19 March*]
  < WHOreliability: := Unreliable > Los expertos< /WHO >
  [*The experts*]
  < WHOreliability: := Reliable > Investigadores del grupo de investigación GRIAL< /WHO >
  [*Researchers of GRIAL research group*]

- **Objectivity**: in a news item, neutrality is a key component. A news item is more likely to be Reliable when information is provided in an objective manner, i.e., it does not positively or negatively influence the reader and does not show the author's stance. In this sense, the Elements of Interest explained in Sect. 3.1.3 provide clues about the objectivity or subjectivity of the content:

  - KEY_EXPRESSIONS: offensive, hopeful, alarming or exhortative messages are a clear sign of unreliability because the author is trying to manipulate the reader and to play with people's emotions. As stated by Zhang et al. (2018), readers can be mislead by the emotional tone of articles and this tone can be found in exaggerated claims or emotionally charged sections, such as expressions of contempt, outrage, spite or disgust.

    < KEY_EXPRESSION > Esta noticia podría salvar el mundo< /KEY_EXPRESSION >
    [*This news could save the world*]
    < KEY_EXPRESSION > Evite que sus amigos y conocidos se enfermen< /KEY_EXPRESSION >

[*Keep your friends and acquaintances from getting sic*]
< KEY_EXPRESSION > Esta gentuza miserable< /KEY_EXPRESSION >
[*This miserable riffraff*]

- QUOTE: this element adds credibility to a news item when it comes from outside experts or organizations and studies (Zhang et al., 2018). However, the author's stance of a QUOTE (indicated with the attribute author_stance and the values Agree, Disagree or Unknown) can also influence the unreliability of a news item. In that case, if the text shows that the author supports or refutes an idea, we will be dealing with a subjective component, as the author will be giving his/her opinion. However, a QUOTE labelled with the value Unknown indicates neutrality since the author will only be reproducing the words of a third party to inform and not to influence the reader.

  < QUOTEauthor_stance: := Disagree >"Nunca se necesitaron los ventiladores, ni la unidad de cuidados intensivos" < /KEY_EXPRESSION >
  ["Ventilators were never needed, nor was the intensive care unit"]
  La IARC califica la acrilamina como < QUOTEauthor_stance: := Unknown >"probable carcinógeno humano"< /QUOTE >
  [IARC classifies acrylamine as a "probable human carcinogen"]

- ORTHOTYPOGRAPHY: this label can also have a negative impact, as spelling mistakes, poor or careless writing style, inadequate punctuation or constant use of capital letters will not be considered a quality news item. Some examples of orthotypography are: whole sentences in capital letters, suspension points in the middle of the text or incomplete, double spaces, many exclamation marks, grammatical errors, spelling mistakes, lack of cohesion, etc.

  < ORTHOTYPOGRAPHY > Con abundantes vaporizaciones< /ORTHOTYPOGRAPHY >
  [*With abundant vaporisations*]

- FIGURE: this can also be used as a reliable characteristic (Rashkin et al., 2017). It is an element that can be easily verified with fact-checking tools.

  Sanidad notifica < FIGURE > 106< /FIGURE > defunciones en las últimas 24 horas
  [*The Spanish health service notifies 106 deaths in the last 24 h*]

- **Title**: unreliable news usually has alarmist, subjective and striking titles. In our annotation proposal, we mark this characteristic with the attribute title_stance, which can be Objective or Subjective. Clickbait title and title representativeness are classified as content indicators by Zhang et al. (2018) and can be misleading or opaque about a topic. In unreliable news, titles tend to be longer and to use more capitalized words (Horne and Adali, 2017) and punctuation marks (especially exclamation marks) and ellipses are usually used (Mottola, 2020).

Table 1 Dataset description (5W1H labels)

Value/label	Reliable %	Unreliable %	Total
WHAT	74.64	25.09	1,100
WHO	84.49	15.37	748
WHEN	78.93	21.07	299
WHERE	94.61	4.79	334
WHY	69.08	30.92	152
HOW	75.74	23.76	202

< TITLEstyle: := Subjectivetitle_stance: := Agree > PRECAUCIÓN: El uso prolongado de la mascarilla produce hipoxia< /TITLE >
[*CAUTION: Prolonged used of the mask causes hypoxia*]
< TITLEstyle: := Subjectivetitle_stance: := Agree >
!‘Grandísimo escándalo! La EMA ve vínculos entre AstraZeneca y los coágulos y trombosis< /TITLE >
[*Huge scandal! The EMA sees links between AstraZeneca and blood clots and thrombosis*]

- **Personal Remarks**: when the author speaks in the first person, tells his/her personal experience or that of someone he/she knows, it is a sign of low credibility, as the author is trying to scare, persuade or make the reader feel closer to the story and thus empathise. These types of personal remarks make the story more subjective and the reader more vulnerable to believe the news item. In fact, we concur with Rashkin et al. (2017), as their results show that "first-person and second-person pronouns are used more in less reliable or deceptive news types".

  < KEY_EXPRESSION > En nuestra opinión< /TITLE >
  [*In our opinion*]
  < KEY_EXPRESSION > Yo lo hago y me ha funcionado muy bien< /KEY_EXPRESSION >
  [*I do it and it works very well*]
  < KEY_EXPRESSION > Mis padres se enfermaban de gripe o faringitis, solo cuando no lo hacían< /KEY_EXPRESSION >
  [*My parents got sick with the flu or pharyngitis, only when they didn't do it*]

- **Lack of Information**: the attribute lack_of_information is highly correlated with unreliability in our annotation scheme. We use it with the 5W1H labels to mark the absence of important data in the text (such as the cause/reason of an event, the subject of the action, etc) as well as to indicate the lack of evidence such as scientific studies or official and verified data. Sometimes, the author states that the information is based on scientific studies without specifying which ones, which provides little credibility. As stated by Mottola

Table 2 Dataset description (Structure and Elements of Interest labels)

Structure and EoI labels	% Appearance
TITLE	100
SUBTITLE	55
LEAD	95
BODY	100
CONCLUSION	62.50
QUOTE	53.75
KEY_EXPRESSION	32.50
FIGURE	63.75
ORTHOTYPOGRAPHY	40

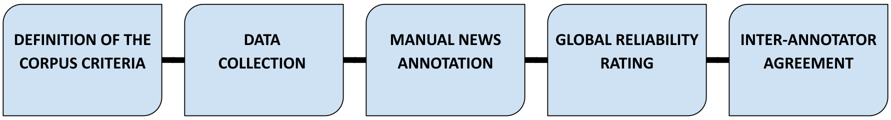

DEFINITION OF THE CORPUS CRITERIA
DATA COLLECTION
MANUAL NEWS ANNOTATION
GLOBAL RELIABILITY RATING
INTER-ANNOTATOR AGREEMENT

Fig. 3 Dataset-building methodology

(2020), the lack of data and sources is another typical characteristic of disinformation, turning news into stories that lack informative content.

< LACK_OF_INFORMATION > Gracias a recientes estudios científicos < /LACK_OF_INFORMATION >
[*Thanks to recent scientific studies*]
< LACK_OF_INFORMATION Según evidencia científica< /LACK_OF_INFORMATION >
[*According to scientific evidence*]

4 RUN dataset

A Reliable and Unreliable News (RUN) dataset in Spanish and focused on health and COVID-19 has been created to test the RUN-AS proposal. The RUN dataset comprises 80 reliable and unreliable news items (36,659 words in total) in Spanish, of which 51 are Reliable and 29 Unreliable, collected from several digital newspapers and manually annotated following the Inverted Pyramid and the 5W1H. Both the annotation of the internal elements (structure and content) and the global reliability of the news item are annotated according to two values: Reliable and Unreliable. Tables 1 and 2 show the total number of labels in the dataset. Both the dataset and the annotation scheme are available at Github repository.[9]

[9] https://github.com/marionieto51/NewsReliabilityAnnotation.

Table 3 Inter-annotator agreement 5W1H labels

Label	IAA
WHAT	0.75
WHO	0.72
WHEN	0.55
WHERE	0.56
WHY	0.37
HOW	0.11
5W1H	0.64

To compile the dataset, three main criteria have been followed: domain, language and traditional news content structure, i.e. Inverted Pyramid. The dataset was delimited to the health and COVID-19 domain and collected during the pandemic, arguably one of the domains that was most exposed to disinformation. The language chosen is Spanish (both from Spain and Latin America) due to the lack of labelled resources in this language. Another criterion taken into account was the traditional news content structure and we only chose news presented in this way, omitting information presented in other formats such as posts, guides, FAQs or social media posts. The methodology for creating the dataset followed five steps and is presented in Fig. 3.

Firstly, the dataset was defined and delimited on the basis of the three aforementioned criteria. Secondly, news was collected both manually and by means of a web crawler. Thirdly, the structure and semantic content of the news items were manually annotated by a linguistic expert annotator following the Inverted Pyramid, the 5W1H and the Elements of Interest levels explained in Sect. 3, and a reliability rating was assigned for each 5W1H label. Fourthly, the global reliability of each news item was assigned by two annotators with knowledge in NLP, taking into account only the plain text, without the labels of the expert annotator. Finally, two inter-annotator agreements were measured to validate both the quality of the dataset and the complexity of the annotation.

4.1 Inter-annotator agreement experiments

In order to assess the quality of the dataset and the complexity of the annotation guideline, two inter-annotator agreements were carried out by using the formula proposed by Névéol et al. (2011): *IAA = number of matches/(number of matches + number of non-matches).*

The first inter-annotator agreement was used to measure the complexity of the annotation labels while the second agreement was performed to provide a global classification to the news in an objective way, without the influence of the internal annotation of an expert linguistic annotator. The IAA in this research measures the success or error between annotators when using the guideline. For that reason, ideas that have not been annotated by both annotators were not compared, since that lack

Table 4 Inter-annotator agreement of reliability when there is agreement in the 5W1H labels

Label	IAA
Reliability of WHAT	0.83
Reliability of WHO	0.87
Reliability of WHEN	0.81
Reliability of WHERE	1.00
Reliability of WHY	1.00
Reliability of HOW	0.00
Reliability of 5W1H	0.85

of coincidence is more related to the degree of subjectivity than to the degree of annotation error.

Other IAA metrics, such as Cohen's *kappa* (Vieira et al., 2010), were discarded as unsuitable given the characteristics of the annotation proposed here. This metric is not relevant for token-level annotation tasks when many tokens are not annotated in the text.

4.1.1 Inter-annotator agreement on annotation labels and reliability

To measure the complexity of the labels related to the journalistic techniques (Inverted Pyramid and essential content), we selected a set of news items not included in the dataset and asked two PhD students with different profiles (linguistics and computer science), both working in the NLP field, to annotate them. The total news items amounted to 1,337 words with reliable and unreliable information. Without previous training, they had to individually annotate news according to the annotation scheme proposed. The objective was to test the difficulty of the scheme and to know whether it is necessary to train the annotators beforehand and how comprehensive the training should be.

The IAA regarding Inverted Pyramid considers a match when the same piece of news content is assigned to the same structure label. The agreement between the annotators was IAA=0.80 in the Inverted Pyramid.

In the case of 5W1H, inter-annotator agreement was measured at three levels of complexity. First, agreement only of 5W1H labels was measured (Table 3). The criteria followed to consider whether or not there is agreement between the annotators was:

- When comparing annotations, we considered that a match occurred when the annotators (A and B) agreed on assigning the same category (WHO, WHAT, WHERE, WHEN, WHY, HOW) to a specific span of elements in the text. A slight difference in length regarding the span of the elements to be annotated is allowed, as long as one string is contained in the other one. For example: "scientists" annotated as WHO by the annotator A and "scientists specialised in biophysics" annotated as WHO by the annotator B.

Table 5 Inter-annotator agreement of 5W1H label and its reliability

Label	IAA
WHAT + Reliability	0.62
WHO + Reliability	0.63
WHEN + Reliability	0.45
WHERE + Reliability	0.56
WHY + Reliability	0.37
HOW + Reliability	0.00
5W1H + Reliability	0.54

- We considered as a non-match those cases where the annotators did not agree to use the same 5W1H label for a span of elements in the text. For example: "scientists" was annotated as WHO by the annotator A and it was annotated as WHERE by the annotator B. There would also be no match if they annotate with a 5W1H different portions of text.

As can be observed from the results, the labels with the highest agreement are WHAT and WHO, whereas the labels WHY and HOW obtain a significantly lower agreement, also due to the semantic complexity of this type of labels.

The second level of IAA calculated was measuring the agreement in the reliability attribute when both annotators agree to annotate the 5W1H label (Table 4). This inter-annotator agreement allows us to evaluate the difficulty of annotating reliability of the 5W1H in isolation and it provides an indication of the suitability of the guideline's criteria used for defining the reliability of the labels. In this case, when both annotators are considering the same 5W1H, there is a match if the reliability value is the same for both annotators. It will be considered a non-match otherwise.

In the case where there is agreement in annotating the 5W1H labels among the annotators, the agreement in annotating the reliability of the label is very high in most cases, which indicates that the defined reliability criteria are suitable. However, in the case of HOW labels, since the subset of news items used was chosen randomly, the amount of HOW annotated labels was not significant. Furthermore, when there was agreement annotating the HOW label there was no agreement in reliability. After analysing the semantic complexity of the parts of text annotated as HOW, these results suggest that it would be necessary to make a more thorough study of how to determine reliability in this type of labels, which may be much less evident than in the case of other tags such as WHO or WHERE.

Thirdly, the last level of inter-annotator agreement presented consists of determining the agreement between the two annotators on both the 5W1H label and its reliability. Both things must coincide to be considered as a match and it will be a non-match otherwise. The agreement between the two annotators at this level is the most complex to achieve (Table 5).

Finally, the expert (author of the annotation guideline) assessed the annotations performed and analysed the inter-annotator agreement results. The conclusion obtained was that annotating semantic elements has a higher level of difficulty and therefore more intense training needs to be provided to annotators for this purpose. To this end, it would be necessary to train annotators with a set of examples and tutorials to show in detail the labels and the attributes. Unlike binary classification annotations (with only two classification values), the annotation proposed is quite accurate and complex, thus it cannot be evaluated without prior training. However, this complexity in turn provides more labelled data to train.

4.1.2 Inter-annotator agreement on global reliability

In order to avoid biases in the overall reliability rating and so that the annotators would not be influenced by the internal labels, two different annotators, with knowledge of NLP, were used from the previous ones in Sect. 4.1.1. They were asked to annotate the global reliability of each document of the RUN Dataset. For this agreement, the annotators had to assign a reliability or unreliability score to the whole RUN dataset but their annotations had to be made using plain text only, without labels, and following the reliability criteria defined in the scheme. The inter-annotator agreement obtained between the two annotators in this task was 0.75 and, although this is considered a fairly high score, it is worth noting that the two annotators analysed the news items they did not agree on in order to exchange ideas and reach a consensus. From this agreement procedure, and with the final assessment of the expert, a global reliability annotation was obtained.

5 Validation of RUN-AS scheme: evaluation framework

Several experiments were conducted to validate our annotation scheme and to support the hypothesis that a fine-grained reliability assessment of multiple semantic elements in a news story can provide an accurate prediction of the global reliability assessment of a news story. The annotated corpus is too small to obtain a robust system for predicting reliable and unreliable news; however, the corpus is appropriate to validate the importance of the features annotated on the task. State-of-the-art (SOTA) ML and DL methods, widely applied in the disinformation classification task, were used to determine whether the information provided by the proposed annotation scheme is feasible to address disinformation detection.

This fine-grained annotation proposal provides linguistic and semantic features that enrich the training process of classification models. From the three annotation levels (Structure, Content, and Elements of Interest), two types of features were extracted: numerical and categorical. In total, 42 different features were extracted per news item.

From the Inverted Pyramid structure level, a total of 7 features were extracted as follows: 5 categorical features (TITLE, SUBTITLE, LEAD, BODY and CONCLUSION) that indicate the presence of these news-structure parts; and, 2 other categorical features extracted from the attributes of the TITLE (stance and style). Concerning the 5W1H content and Elements of Interest levels, there is a total of 35 numerical features that refer to the number of labels for each one. As for the

5W1H content level, 6 features were extracted related to each 5W1H (WHAT, WHO, WHERE, WHEN, WHY and HOW). For each 5W1H label, the number of attributes of type Reliable/Unreliable was counted (12 features), as well as the number of the attributes of type lack_of_ information (6 features), the attribute of type role (3 features), the attribute of type main_event (1 feature). Regarding the level of Elements of Interest, a total of 4 numerical features were extracted (FIGURE, KEY_EXPRESSION, ORTHOTYPOGRAPHY and QUOTE), as well as the number of attributes of type author_stance (3 features).

A simplified example of the numerical and categorical features extracted from the TITLE and LEAD of a news piece is presented next.[10]

```
{
     TITLE_style : Objective,
     TITLE_stance : Agree,
     TITLE_WHAT_Reliable : 0,
     TITLE_WHAT_Unreliable : 1,
     TITLE_WHO_Reliable : 0,
     TITLE_WHO_Unreliable : 1,
     TITLE_WHEN_Reliable : 0,
     TITLE_WHEN_Unreliable : 1,
     #...
     LEAD_WHAT_Reliable : 2,
     LEAD_WHAT_Unreliable : 2,
     LEAD_WHO_Reliable : 0,
     LEAD_WHO_Unreliable : 1,
     LEAD_WHEN_Reliable : 0,
     LEAD_WHEN_Unreliable : 3,
     #...
}
```

The same type of features will be generated from the other parts of the structure of the document. Each feature indicates the number of 5W1H components with a specific label and reliability attribute that appear in each part of the news. For example, `LEAD_WHAT_Reliable: 2` indicates that the `LEAD` contains two `WHAT` items annotated with a `Reliable` value. The model is trained to predict the overall document reliability label based on these numerical and categorical features.

[10] Only some of the features are shown to exemplify the generation of these features.

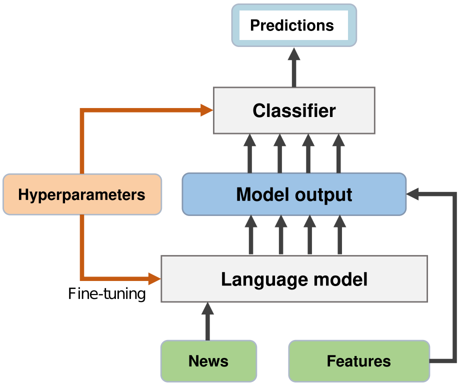


Fig. 4 Classification architecture with external features. Update of Sepúlveda-Torres et al. (2021)

5.1 Experiments

To validate the RUN-AS proposal, the experiment section has a twofold objective. First, validation of the proposal by means of ML and DL (pre-trained transformer models) methods used in the disinformation tasks, with the objective of enhancing the performance of the task when training the models with the dataset annotated with RUN-AS. And second, an analysis of the influence of the different features used in the annotation scheme to determine which of these elements of the scheme are more decisive than others when classifying the reliability. To that end, the following three experiments were carried out:

1. **ML performance**: The following ML classification algorithms are used to create baseline systems and train the classifiers: Support Vector Machines (SVM), Random Forest, Logistic Regression, Decision Tree, Multi-layer Perceptron (MLP), Adaptive Boosting (AdaBoost) and Gaussian Naive Bayes (GaussianNB). Two configurations of the aforementioned algorithms are used.

    - *Baseline model:* Encoding of news texts by using TF-IDF type vectors.
    - *Model with RUN-AS features*: Concatenation of the TF-IDF vectors with the 42 features obtained from the annotation.

    This experiment was implemented using *scikit-learn*.[11] It can be replicated at the Colab[12] notebook.
2. **DL performance (Pre-trained transformer model):** In the context of DL models it is common to find the use of huge datasets. However, some DL models take advantage of the transfer learning technique, whereby knowledge is transferred from a previously trained model in a general task to specific tasks by using a lower cardinality dataset. This avoids the effort of starting the learning from scratch (Pan and Yang, 2010). Subsequently, these pretrained models are used to

[11] https://scikit-learn.org/stable.

[12] https://github.com/rsepulveda911112/ML_RUN_Dataset.

perform a fine adjustment that consists of retraining the model on another dataset, to readjust the weights of the network in the specific domain (Tajbakhsh et al., 2016). Taking into account the preliminary annotation made, we consider that by using a model based on transfer learning we can avoid learning problems with the dataset.

In this context the Beto[13] language model based on transformer architecture (for more detail consult Canete et al. (2020)) was used to create two classifier models. Both classifier models consist of fine-tuning the model by using the annotated dataset and are composed of two main components: a language model (BETO) and a classification neural network. Figure 4 shows the architecture of the classification used. The following hyperparameters were used: maximum sequence length of 512, batch size of 2, training rate of 2e-5, and training performed for 3 epochs.

- *Baseline model:* The first is a baseline system that used the news as input to the language model (BETO).
- *Model with RUN-AS features*: The second used the architecture propose by Sepúlveda-Torres et al. (2021), which modified the BETO baselines to include external features. Both the text and the 42 features were used as input. Features are concatenated with the output of the BETO language model to feed the input to the classification neural network.

To create the classifiers, the *Simple Transformers library*[14] was used, which creates a wrapper around *HuggingFace's Transformers library* for using Transformer models (Wolf et al., 2019). These experiments can be reproduced on the repository.[15]

3. **Analysis of the features' influence:** The logistic regression algorithm is used to evaluate the influence of the features extracted from the annotated labels to classify the news as Reliable or Unreliable. Furthermore, an ablation study was performed.

5.2 Cross-validation strategy

The cross-validation strategy was performed in all experiments. This is a statistical technique that involves partitioning the data into subsets, training the data on a subset, and using the other subset to evaluate the model's performance. Cross-validation enables all available data to be used for training and testing (Bergmeir and Benítez, 2012). This technique is used to determine how the results of a machine learning model could be generalized to new, unseen data. In these experiments, k-fold cross-validation with k = 5 is used, where 80% of each subset has been used for training and 20% for testing.

[13] https://github.com/dccuchile/beto .

[14] https://simpletransformers.ai/.

[15] https://github.com/rsepulveda911112/BETO_RUN_AS.git.

Table 6 Experiments results using ML and DL methods

Experiments	Baseline model (TF-IDF)		Model with RUN-AS features	
	Acc	F_1m	Acc	F_1m
SVM	0.662	0.395	0.937	0.925
Random Forest	0.75	0.639	0.912	0.898
Logistic Regression	0.65	0.392	0.912	0.875
Decision Tree	0.737	0.683	**0.95**	**0.948**
MLP	0.712	0.57	0.925	0.912
AdaBoost	**0.787**	**0.748**	**0.95**	0.945
GaussianNB	0.612	0.456	0.687	0.57
	Baseline model		Model with RUN-AS features	
BETO	**0.85**	**0.80**	**0.887**	**0.854**

The best results are marked in bold

In order to evaluate the proposal, the commonly used NLP measures (accuracy and macro-averaged F_1 –F_1m–) are used.

6 Validation of RUN-AS scheme: results and discussion

This section presents the results obtained in each of the experiments described in Sect. 5 and a discussion of those results.

6.1 ML and DL performance results

Table 6 presents the performance of experiments 1 and 2 explained in section 5.

As can be concluded from the results in the table, all the models that used features obtained by annotation significantly outperform the proposed baselines. The best results are attained with Decision Tree using RUN-AS annotation, obtaining a 0.948 of macro F_1 (F_1m). It is noteworthy that when using the whole document annotated without external features (baselines) the best F_1m value is obtained by the DL approach (BETO) with 0.80 F_1m, followed by AdaBoost with 0.748 F_1m. However, for the rest of the approaches, the results obtained by only using the document are very poor. All approaches are significantly improved by using the information provided by the annotation labels of the RUN-AS scheme. The worse results are obtained with Logistic regression. However, applying the annotation scheme to learn the model in this case is able to increase the F_1m by 0.483 points. Therefore, these results validate the main hypothesis presented in this research, i.e., that individual 5W1H components reliability are a good predictor of overall news story reliability.

Table 7 Individual weights associated to each of the numerical features extracted from the training dataset. Negative weights count towards the Unreliable class and positive weights towards the Reliable class

Annotation	Attribute	Value	Weight
WHAT	reliability	Unreliable	−0.7108
WHAT	main_event		−0.7071
WHAT	lack_of_information	Yes	−0.6961
WHY	reliability	Unreliable	−0.4729
WHO	reliability	Reliable	−0.4013
TITLE	style	Subjective	−0.3382
WHO	role	Both	−0.3354
HOW	reliability	Reliable	−0.3176
QUOTE	author_stance	Agree	−0.3172
LEAD			−0.2687
QUOTE			−0.2427
TITLE	title_stance	Agree	−0.2154
WHAT			−0.1857
WHEN	reliability	Unreliable	−0.1302
SUBTITLE			−0.1064
WHY	lack_of_information	Yes	−0.0979
WHERE	lack_of_information	Yes	0.0907
WHEN	lack_of_information	Yes	0.0888
TITLE	title_stance	Unrelated	−0.015
WHEN			0.0188
WHO	lack_of_information	Yes	0.0219
WHERE	reliability	Reliable	0.025
CONCLUSION			0.0255
HOW	lack_of_information	Yes	0.0575
QUOTE	author_stance	Disagree	0.0578
FIGURE			0.0751
ORTHOTYPOGRAPHY			0.1987
WHO	role	Subject	0.1988
HOW			0.2297
TITLE	title_stance	Disagree	0.2304
WHO			0.2574
TITLE	style	Objective	0.3382
WHERE	reliability	Unreliable	0.3447
WHERE			0.3499
WHAT	reliability	Reliable	0.4117
WHO	role	Target	0.4406
WHY	reliability	Reliable	0.4793
HOW	reliability	Unreliable	0.5473
WHO	reliability	Unreliable	0.5783

Table 8 Ablation study using the Decision Tree algorithm

Experiments	Model with RUN-AS features	
	Acc	F_1m
Without Inverted Pyramid structure level	0.962	0.961
Without 5W1H content	**0.925**	**0.899**
Without Elements of Interest levels	0.937	0.934

The worst results are marked in bold

Analyzing the results comparing the DL and ML approaches, it can be seen that without the use of the annotated features, the DL approach obtains significantly higher results than those obtained by the ML approaches, demonstrating the power of the transformer architecture to encode text and find implicit features in them. However, when using the annotated features most ML approaches outperform the results achieved by BETO. A possible explanation to explain this behaviour could be that the architecture used to introduce the annotated features with the BETO model has not been able to accurately exploit these features as the classical machine learning algorithms used have.

6.2 Analysis of features' influence results

The high test accuracy obtained in the different experiments performed does support the hypothesis that the fine-grained annotations provided in RUN-AS are highly correlated with the overall news reliability. Furthermore, analysing the individual weights attributed to each of the extracted features provides a deeper understanding of the importance of each annotated element. Logistic regression is used since it is a classifier that enables individual weights, associated to each of the features extracted, to be obtained from the annotation scheme.

Table 7 summarises the learned weights for the most interesting features extracted from the annotated documents. Negative weights indicate that the presence of the corresponding features is correlated with an `Unreliable` label for the overall news item, and positive weights with a `Reliable` label. The closer to zero, the less representative are the features for classification, and the further from zero, the more representative are the features for the classification model.

The distribution of weights is aligned with expected behavior. Annotations of type `WHAT`, `WHY` and `WHO` have a strongly correlated weight with an `Unreliable` global score, especially in the case of `WHAT`, when the `WHAT` is Unreliable, contains main_event or lack_of_information.

Likewise, the `style` attribute of the `TITLE` element is strongly correlated in a meaningful sense, with subjective TITLES indicating a higher probability of an `Unreliable` news item (negative weight) and vice versa.

In addition, to determine the influence of the different features used in our experiments, an ablation study was performed, using the approach with best performance (Decision Tree). The ablation study consisted of performing three different experiments, each time removing one specific group of features, with the aim of gaining better insights on how each of the three annotation levels (Structure, Content, and Elements of Interest) contributes to the proposal. The results are presented in Table 8.

An analysis of the ablation study results indicates that the most influential group of features for the model are those related to its content (those that take semantic aspects into account). On the other hand, the experiment that does not use the features from the Inverted Pyramid level (*Without Inverted Pyramid structure level experiment*) obtains better results than the experiment that uses all the features. This preliminary result suggests that the features which indicate the presence of the parts of the inverted pyramid are not relevant for predicting the reliability of a news item. In future research, carrying out experiments that involve more news items is likely confirm this discovery with greater certainty.

7 Automatic detection of reliability of 5W1H components

The experiments presented in the previous sections (Sects. 5 and 6) allow to demonstrate within a theoretical framework that the features annotated by RUN-AS allow to improve the reliability detection task using a relatively small corpus and a limited computational capacity. Nevertheless, we are aware of the current limitations of the proposal since a manual annotation of the features is used and in a real environment this information could not be available beforehand, limiting the performance of the system. Considering these limitations, in a real environment it would be necessary to determine all the labels and their reliability automatically and as accurately as possible, in order to achieve a performance similar to the theoretical framework proposed here. Applying existing tools in NLP, the final goal would be to automatically detect these features in plain text as efficiently as possible.

In order to demonstrate the real feasibility of the proposal, as a first step, we trained a model to detect the reliability of the 5W1H elements. An experiment was performed on the dataset annotated exclusively with the structure and content labels.

To carry out the reliability classification of the 5W1H, the BETO model was used with a similar configuration to the experiments in section 5.1. In this case, the text of the 5W1H labels were used as input to the BETO model. The same cross-validation strategy was performed to train and validate, obtaining 0.9 accuracy and 0.73 F_1m. These good results in the automatic detection of the reliability of the labels corroborate the feasibility of the proposal in a future fully-automatic pipeline, from plain text to the annotation of the reliability of the parts and of the overall news.

The next step would be the automatic detection of the journalistic structure and content elements of a news item. This implies the use of extra NLP resources to automatically annotate these elements, which is beyond the scope of the present research. Even so, we are working on this further work which will also allow us to integrate it into a semi-automatic annotation system that will enable the generation of these complex datasets in a faster and more efficient way.

8 Conclusions and future work

The novelty of this work lies in the development of RUN-AS, a fine-grained annotation scheme based on journalistic techniques that classify news into Reliable or Unreliable. This annotation proposal was tested by using ML and DL experiments in a Spanish news dataset called RUN, created ad hoc. Furthermore, to assess the complexity of the structure and content annotation as well as annotate the global reliability of the news item, we have measured annotations of four annotators with knowledge of NLP via two inter-annotator agreements. This was done to mitigate biases that might be included from the expert linguistic annotator.

Experiments conducted have shown that the individual reliability of each of the elements annotated contributes to know the overall reliability of a news item with a 0.948 F_1m performance. Furthermore, an analysis of which features have a greater influence in classifying news as Reliable or Unreliable was performed. Therefore, the experiments presented here supports the hypothesis that a fine-grained reliability assessment of multiple semantic elements in a news story can provide an accurate estimate of a global reliability score.

A classification into Reliable or Unreliable could support fact-checking techniques, representing a previous stage that generates a reliability report for quickly checking a news item. This could be employed before fact-checking or it could be used as a tool to support writers and journalists to enhance the accuracy of their work. This annotation can be complementary to other lines of research, such as fact-checking or contradiction detection, as it provides useful information on a first level of a text-only annotation.

Our proposal is designed to annotate the way in which news is written and communicated, the style, the structure of the story, the tone, the evidence, the neutrality or the way in which information is provided. There are key characteristics that distinguish Reliable from Unreliable news. Furthermore, the RUN-AS follows two important premises in journalism: the Inverted Pyramid and the 5W1H. These two concepts are the basis of our annotation, as they enable the annotation of all the structural parts and the semantic elements of a news items.

As future work, we will focus on extending our dataset following the annotation scheme presented in this study. We are already working on an assisted annotation proposal that combines both manual and automatic approaches, using active learning and human-in-the-loop methodologies. This semi-automatic system will reduce the time and the effort spent on compilation and annotation tasks, allowing a complex annotation with high accuracy. In addition, we will aim to create an automated model to annotate both the Inverted Pyramid structure and the 5W1H content with their reliability from plain news items.

Author contributions AB-J: conceptualization, methodology, investigation, resources, data curation, writing—original draft, writing—review/editing, visualization, funding acquisition. RS-T: conceptualization, methodology, software, validation, formal analysis, investigation, resources, writing - original draft, writing—review/editing, visualization. ES: conceptualization, methodology, investigation, resources, writing - original draft, writing—review/editing, supervision, project administration, funding acquisition, visualization. PM-B: conceptualization, methodology, investigation, writing—review/editing,

supervision, project administration, funding acquisition. MN-P: methodology, resources, software, formal analysis, validation.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research work is funded by MCIN/AEI/ 10.13039/501100011033 and, as appropriate, by “ERDF A way of making Europe”, by the “European Union” or by the “European Union NextGenerationEU/ PRTR” through the project TRIVIAL: Technological Resources for Intelligent VIral AnaLysis through NLP (PID2021-122263OB-C22) and the project SOCIALTRUST: Assessing trustworthiness in digital media (PDC2022-133146-C22). Also funded by Generalitat Valenciana through the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/ 2021/21), and the grant ACIF/2020/177.

Data availability The experiments can be reproduced at the Colab notebook https://github.com/rsepulveda911112/ML_RUN_Dataset and on the repository https://github.com/rsepulveda911112/BETO_RUN_AS.git. Both the dataset and the annotation guideline created for this research are available at Github respository: https://github.com/marionieto51/NewsReliabilityAnnotation.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Assaf, R., & Saheb, M. (2021). Dataset for arabic fake news. In *2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), IEEE* (pp. 1–4).

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr, 2020*2020, 1-10

Chakma, K., & Das, A. (2018). A 5w1h based annotation scheme for semantic role labeling of english tweets. *Computación y Sistemas, 22*(3), 747–755.

Chakma, K., Swamy, S. D., Das, A., & Debbarma, S. (2020). 5w1h-based semantic segmentation of tweets for event detection using bert. In I*nternational Conference on Machine Learning, Image Processing, Network Security and Data Sciences* (pp 57–72). Springer

DeAngelo, T. I., & Yegiyan, N. S. (2019). Looking for efficiency: How online news structure and emotional tone influence processing time and memory. *Journalism & Mass Communication Quarterly, 96*(2), 385–405.

Ferreira, W., & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp 1163–1168. https://doi.org/10.18653/v1/N16-1138, https://aclanthology.org/N16-1138.

Figueira, Á., & Oliveira, L. (2017). The current state of fake news: Challenges and opportunities. *Procedia Computer Science, 121*, 817–825.

Giansiracusa, N. (2021). *How Algorithms Create and Prevent Fake News*. Springer.

Gruppi, M., Horne, B. D., & Adali, S. (2018). An exploration of unreliable news classification in brazil and the us. arXiv preprint arXiv:1806.02875.

Hamborg, F., Breitinger, C., Schubotz, M., Lachnit, S., & Gipp, B. (2018). Extraction of main event descriptors from news articles by answering the journalistic five w and one h questions. In*Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, (pp. 339–340).

Hordofa, B. A. (2020). Event extraction and representation model from news articles. *16*, 1–8.

Horne, B., & Adali, S. (2017). 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*, (pp. 759–766).

Hou, L., Li, J., Wang, Z., Tang, J., Zhang, P., Yang, R., & Zheng, Q. (2015). Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems, 76*, 17–29.

Khodra, M. L. (2015). Event extraction on indonesian news article using multiclass categorization. In 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), IEEE, pp 1–5.

Mottola, S. (2020). Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español. *Discurso & Sociedad, 3*, 683–706.

Névéol, A., Doğan, R. I., & Lu, Z. (2011). Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction. *Journal of biomedical informatics, 44*(2), 310–318.

Norambuena, B., Horning, M., & Mitra, T. (2020). Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. In *Computational Journalism Symposium*.

Paka, W. S., Bansal, R., Kaushik, A., et al. (2021). Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing, 107*(107), 393.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation* (pp. 21–29). Springer.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015).*The development and psychometric properties of liwc2015*. Tech. rep.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint* arXiv:1708.07104.

Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., & Escobar, J. J. M. (2019). Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems, 36*(5), 4869–4876.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). 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*, (pp. 2931–2937).

Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., & Palomar, M. (2020). Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications, 141*(112), 943.

Sepúlveda-Torres, R., & Saquete Boró, E. (2021). Gplsi team at checkthat! 2021: Fine-tuning beto and roberta. CEUR.

Shahi, G. K., Struß, J. M., & Mandl, T. (2021). Overview of the clef-2021 checkthat! lab task 3 on fake news detection. Working Notes of CLEF.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint* arXiv:1707.0759296, 104.

Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. In *Disinformation, Misinformation, and Fake News in Social Media*, (pp. 1–19).

Silva, R. M., Santos, R. L., Almeida, T. A., & Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications, 146*(113), 199.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102–107).

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging, 35*(5), 1299–1312.

Thomson, E. A., White, P. R., & Kitley, P. (2008). objectivity and hard news reporting across cultures: Comparing the news report in english, french, japanese and indonesian journalism. *Journalism studies, 9*(2), 212–228.

Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010). Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems , IEEE* (pp. 1–8).

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In*Proceedings of the ACL 2014 workshop on language technologies and computational social science* (pp. 18–22).

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint* arXiv:1705.00648.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). *Huggingface's transformers: State-of-the-art natural language processing*. arXiv:1910.03771.

Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., ... & Mina, A. X. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference* (pp 603–612).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alba Bonet-Jover[1] · **Robiert Sepúlveda-Torres**[1] · **Estela Saquete**[1] · **Patricio Martínez-Barco**[1] · **Mario Nieto-Pérez**[1]

✉ Alba Bonet-Jover
alba.bonet@dlsi.ua.es

Robiert Sepúlveda-Torres
rsepulveda@dlsi.ua.es

Estela Saquete
stela@dlsi.ua.es

Patricio Martínez-Barco
patricio@dlsi.ua.es

Mario Nieto-Pérez
mnieto@dlsi.ua.es

[1] Department of Software and Computing Systems, University of Alicante, carretera San Vicente s/n, San Vicente del Raspeig 03690, Alicante, Spain