Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 15-26 recibido 12-12-2022 revisado 09-02-2023 aceptado 15-02-2023

Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation

Anotando la confiabilidad para mejorar la tarea de detección de desinformación: esquema de anotación, recurso y evaluación

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, Patricio Martínez Barco
Department of Software and Computing Systems, University of Alicante, Spain
{alba.bonet, rsepulveda, stela, patricio}@dlsi.ua.es

Abstract: Disinformation is a critical problem in our society. The COVID-19 pandemic and the Russia-Ukraine war have been key events for the spreading of fake news. Assuming that fake news mixes reliable and unreliable information, we propose RUN-AS (Reliable and Unreliable Annotation Scheme), a fine-grained annotation scheme that labels the structural parts and essential content elements of a news item to enable their classification into Reliable and Unreliable. This type of annotation will be used for training systems to automatically classify the reliability of a news item. To this end, RUN dataset in Spanish was built and annotated with RUN-AS. A set of experiments were conducted to validate the annotation scheme. The experiments evidence the validity of the annotation scheme proposed, obtaining the best F_1m, i.e., 0.948.
Keywords: Natural Language Processing, Annotation Guideline, Dataset Annotation, Reliability Detection, Disinformation Detection.

Resumen: La desinformación es un problema crítico en nuestra sociedad. La pandemia de covid-19 y la guerra entre Rusia y Ucrania han sido escenarios clave para la difusión de noticias falsas. Partiendo de la base de que las noticias falsas mezclan información confiable y no confiable, proponemos RUN-AS (*Reliable and Unreliable Annotation Scheme*), un esquema de anotación de grano fino que etiqueta las partes estructurales y los elementos de contenido esenciales de una noticia y permite clasificarlos en Confiable y No confiable. Esta anotación será usada en el entrenamiento de sistemas para la clasificación automática de la confiabilidad de una noticia. Para ello, se construyó el corpus RUN en español y se anotó con RUN-AS. Se llevó a cabo un conjunto de experimentos para validar el esquema de anotación. Los experimentos evidencian la validez del esquema de anotación propuesto, obteniendo el mejor F_1m 0,948.
Palabras clave: Procesamiento Lenguaje Natural, Guía Anotación, Anotación Corpus, Detección Confiabilidad, Detección Desinformación.

1 Introduction

The disinformation problem is critical for today's society. Disinformation is fake or inaccurate information that is intentionally spread to mislead or deceive (Shu et al., 2020). Fake news is one of the most widespread phenomena of disinformation and, as defined by Zhou and Zafarani (2020), fake news is intentionally false information created by journalists and non-journalists that broadly includes articles, claims, statements, speeches, and posts, among other types of information, related to public figures and organizations.

The Internet has made it possible to be continuously informed, driving an almost instant dissemination of unverified news, as anyone can share and access information at no cost. A complex mix of cognitive, social and algorithmic biases makes us more vulnerable to believing and being manipulated by online disinformation (Shao et al., 2017). Algorithms make possible the exponential spread of fake news, but they can

ISSN 1135-5948 DOI 10.26342/2023-70-1

also be deployed to mitigate their propagation (Giansiracusa, 2021). Therefore, the facilitator of the disinformation problem, the algorithm, can be used to combat 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). 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 (Saquete et al., 2020).

The concepts of reliability and veracity are closely related, as fake news includes both reliable and unreliable information. In the literature, the term veracity is usually used in tasks where information is contrasted and verified (Vosoughi, Roy, and Aral, 2018), whereas the concept of reliability is often used in methods that investigate the credibility of the source of the news item (Zhou and Zafarani, 2020). In this research, as we tackle the problem by using the news item and not external knowledge, an absolute judgment on the veracity of a text is not possible. Instead of focusing on the veracity concept, we deal with the concept of reliability by following a style-based method that enables the detection of unreliable elements in a text through linguistic indicators as they mark the inaccuracy or subjectivity of the information provided (Zhang et al., 2018). To address this task computationally, annotated datasets are required (Stenetorp et al., 2012). This is a costly, slow and time-consuming task and therefore, labelled corpora are scarce, especially in languages other than English, such as Spanish.

The novelty of our proposal is the design of an innovative semantic annotation scheme that focuses on classifying news as Reliable or Unreliable from a linguistic perspective and without external knowledge. This annotation scheme will be beneficial to future disinformation detection tasks. 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 their reliability. Furthermore, reliability criteria followed in the annotation process is clearly defined in Section 3.2. Following the proposed annotation guideline, a new dataset (RUN dataset) is created and used to validate the RUN-AS scheme under an evaluation framework. Furthermore, the language used for the annotation scheme and the dataset is Spanish due to the lack of resources in languages other than English.

This paper is structured as follows: Section 2 presents the background; Section 3 describes the annotation scheme proposed; Section 4 introduces the dataset created to test our proposal and two inter-annotator agreements to avoid bias in assessing news; Section 5 presents several experiments that validate our annotation scheme; Section 6 summarises the results and discussion; and finally, Section 7 presents the conclusions of this research and future work.

2 Related Work

This section presents relevant literature regarding state-of-the-art (SOTA) disinformation datasets, work regarding journalistic techniques applied in our proposal and finally, literature regarding research about linguistic characteristics of news in order to detect disinformation.

2.1 Annotated corpora for disinformation detection

Several datasets have been released for disinformation detection. LIAR dataset (Wang, 2017) comprises 12,836 real-world short statements classified in a scale of six fine-grained labels (pants-fire, false, barely-true, half-true, mostly-true and true). EMERGENT dataset (Ferreira and Vlachos, 2016) contains 300 claims and 2,595 associated news articles. This dataset 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). Ferreira and Vlachos (2014) also released a fake news detection dataset comprising 221 statements annotated with a five-label-tag classification: true, mostlytrue, halftrue, mostlyfalse and false. Pérez-Rosas et al. (2017) introduced two new datasets for fake news detection covering several domains and linguistic differences between legitimate and fake news articles. The CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection (Shahi, Struß, and Mandl,

2021) is a lab that focuses on evaluating automatic detection of the news story's veracity, classified as true, partially true, false, or other. The dataset consists of 900 news articles, leaving 354 articles for testing.

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 (Patwa et al., 2021) and a large COVID-19 Twitter Fake News dataset (CTF) (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 fake news dataset 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). In Portuguese, a dataset of labeled true and fake news called the Fake.Br corpus was presented (Silva et al., 2020). It is composed of 7,200 news (fake and legitimate). 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. Regarding datasets that annotate reliability, Gruppi et al. (2018) constructed two datasets of political news articles from United States sources (1,997 reliable, 794 unreliable and 50 satire) and Brazilian sources (4,698 reliable, 755 unreliable and 58 satire). For each article, they assigned a class reliable (R), unreliable (U) or satire (S) based on the source from which the article was collected.

To the authors' knowledge, most current datasets classify and annotate news with a single global veracity value. 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 two well-known journalistic concepts such as the Inverted Pyramid and the 5W1H[1], this subsection focuses on presenting some corpora that also use them. 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 applied to a corpus of 3,000 tweets related to the US elections of 2016. Khodra (2015) introduces a new 5W1H corpus of 90 Indonesian news articles to train event extraction. They were obtained from popular news websites and annotated following the 5W1H concept and extracting the event information of the news item.

The novelty of our annotation compared to the state of the art lies in the annotation of the 5W1H of all parts of a news item, permitting more in-depth analysis of the whole news article.

2.3 Research focused on linguistic features to detect disinformation

This subsection presents the research relevant to analysing linguistic features in news to determine reliability.

Zhang et al. (2018) present a set of content and context indicators for article reliability. Regarding the content indicators, which are the ones that are of interest to our research, the following are considered: title representativeness; clickbait title; quotes from outside experts; citation of organizations and studies; calibration of confidence; logical fallacies; and, tone and inference. Their dataset consists 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 news. In this study, three content-based features categories are analysed: stylistic, complexity, and psychological. Horne and Adali (2017) conclude that there is a notable difference in titles

[1]Referring to: who, what, where, when, why, how.

and content between fake and real news in terms of length, punctuation, quotations, lexical features or capitalised words. 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. Through this linguistic analysis, it is shown that there are several characteristics that fake news share related to headlines, punctuation, capital letters, lack of data or emotional aspects.

Our proposal makes a threefold contribution to disinformation detection. Firstly, a proposal of reliability classification instead of veracity, considering linguistic features, without external knowledge. Secondly, instead of exclusively annotating the entire article with a single global classification value, we also annotate all the structural parts and essential content of a news item in line with the 5W1H and Inverted Pyramid. Thirdly, this fine-grained annotation produces a quality resource in Spanish.

3 RUN-AS annotation scheme

3.1 Annotation labels

The goal of this annotation proposal is to support disinformation detection by analysing news on the basis of a purely textual and linguistic analysis and thereby explore how a news item's structure and wording influence its reliability. RUN-AS (Reliable and Unreliable News Annotation Scheme)[2] is a fine-grained annotation scheme 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 labels (Inverted Pyramid), Content labels (5W1H) and Elements of Interest labels (EoI). Structure labels contain content and elements of interest labels within them. Content and EoI labels can be overlapped.

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, White, and Kitley, 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 (DeAngelo and Yegiyan, 2019). The five structure labels of our proposal are TITLE, SUBTITLE, LEAD, BODY and CONCLUSION. Depending on the source, 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. The definition of the structure labels is:

TITLE: headline of the news item. This label has two possible attributes. The attribute **title_stance** 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). The attribute **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.

SUBTITLE: sentence completing the information of the TITLE.

LEAD: first paragraph presenting the essential information of the news item. It develops and usually repeats the idea presented in the title.

BODY: set of paragraphs developing the story and presenting in detail all the information of the news.

CONCLUSION: last sentence or paragraph summarising the content of the news article. It is not always present.

3.1.2 Content labels

The other technique used 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 un-

[2] Available at `http://bit.ly/3T4XMzn`

derstand 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 reliability than a news item that does not communicate the information in such a precise way. All the 5W1H elements are annotated as Reliable or Unreliable depending on their level of accuracy and objectivity (reliability attribute explained next).

WHAT: facts, circumstances, actions. Example: *los contagios de coronavirus se disparan* (coronavirus infections skyrocket).

WHO: subject, entity. Example: *la Agencia Europea del Medicamento* (European Medicines Agency).

WHEN: time, moment. Example: *el 20 de diciembre* (on 20 December).

WHERE: place, location. Example: *en España* (in Spain).

WHY: cause, reason. Example: *a causa de la muerte* (due to the death).

HOW: manner, method. Example: *con abundante agua* (with abundant water).

The 5W1H labels have the following attributes:

reliability is the main attribute of our annotation and allows to classify each element as well as the global news item with the values R (Reliable) or U (Unreliable), depending on the level of accuracy, objectivity, and the linguistic characteristics.

lack_of_information is used to indicate evidence is missing. This attribute has a single value (Yes). It is not indicate otherwise.

role is the attribute used with the WHO label only. It indicates the role played by the WHO entity in the event. It 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).

main_event is only used with the WHAT label when the WHAT indicates the main event(s) of the story. It is possible to find several events (each one with its own 5W1H), but one is considered the main event.

3.1.3 Elements of Interest labels

The following Elements of Interest labels enable the annotation of textual information that could distinguish Unreliable from Reliable news:

QUOTE: label that marks the presence of quotes in the news item. It has the attribute **author_stance** that serves to annotate the author's stance regarding the QUOTE content. It has three values: Disagree (to express its disagreement towards the idea), Agree (to share its agreement) or Unknown (neutral stance). For example: *el experto niega que "el limón cura el cáncer"* (the expert denies that "lemon cures cancer") is a QUOTE with Disagree author_stance.

KEY_EXPRESSIONS: label containing phraseology that urges readers to share the information or that expresses emotions or economic purposes. For example: v*amos a salvar vidas compartiendo esta gran información* (let's save lives by sharing this important information)

FIGURE: numerical values in a news item.

ORTHOTYPOGRAPHY: label annotating poor writing and text with grammatical, spelling or formatting mistakes.

Figure 1 presents the specification of the three types of levels of the RUN-AS annotation scheme together with the attributes for each label, and the possible values for each attribute.

3.2 Reliability criteria

This work focuses on assigning a reliability value to the essential content labels described in our annotation scheme.

There are textual and linguistic features that enable the detection of the reliability of a news item and of each part of the news item, permitting an assessment of the news item's overall reliability. The criteria used when classifying the reliability consider accuracy and neutrality of the content relies on the state-of-the-art research presented in Section 2.3.

3.2.1 Accuracy

Accuracy is one of the key factors in determining the reliability of information. In our reliability modeling we have considered the following clues:

Vagueness and 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. For example, a reliable WHEN is: *el*

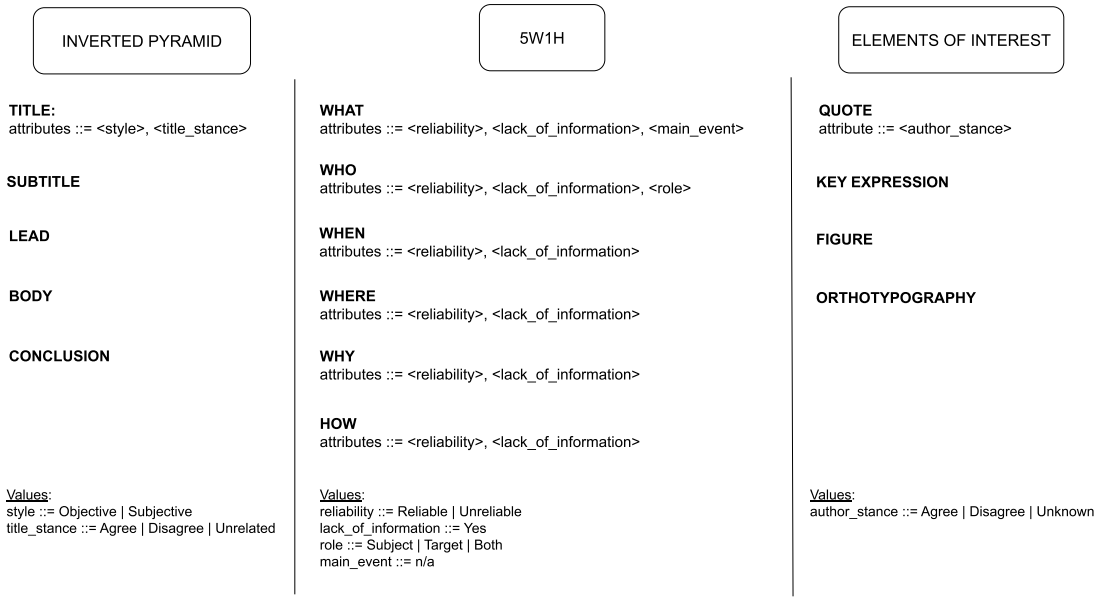

Figure 1: RUN-AS annotation scheme.

viernes 19 de marzo (on Friday 19 March) whereas *hace mucho tiempo* (a long time ago) lacks of accuracy. The existence of vagueness or ambiguity will be annotated with an **U**nreliable value associated with the corresponding 5W1H label.

On the contrary, the presence of figures, annotated with the **FIGURE** EoI label indicates accurate information that can be easily fact-checked with external sources, thus denoting reliability, for instance se han administrado {FIGURE: *6.000.000*} *dosis de vacunas* (6,000,000 doses of vaccine have been administered).

Lack of information. The presence of the this attribute can be considered a signal of unreliability. It appears 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, 2020), the lack of data and sources is another typical characteristic of disinformation, turning news into stories that lack informative content. For example, a WHAT label with lack_of_information attribute is: *según algunos científicos* (according to some scientists).

Typos. When the **ORTHOTYPOGRAPHY** label is annotated, it has a negative reliability 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 sentences; double spaces; many exclamation marks; grammatical errors; spelling mistakes; lack of cohesion; etc. For instance, *aquí en nuestro Pays* (here in our "Countri") is annotated with the ORTHOTYPOGRAPHY label.

3.2.2 Neutrality

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 and does not show the author's stance. Hints about text neutrality (or lack thereof) are considered in the RUN-AS schema as follows:

Personal Remarks and Emotional Messages. 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 (Rashkin et al., 2017). Futhermore, offensive, hopeful, alarming or exhortative messages are a clear sign of unreliability because the author is try-

ing to manipulate the reader and to play with people's emotions (Zhang et al., 2018).

Through the labeling of **KEY_EXPRESSIONS** we can represent this kind of non-neutral information.

Some examples of KEY_EXPRESSIONS regarding this issue are: *yo lo hago y funciona* (I do it and it works) or *evite que sus amigos y conocidos se enfermen* (keep your friends and acquaintances from getting sick).

Quotes and author stance. The presence of **QUOTE** labels add neutrality to a news item since it indicates that the information comes from an external source (Zhang et al., 2018). However, when the author is clearly in favor or against the quote, an important hint of subjectivity is introduced. Thus, labeling QUOTE with attribute **author_stance=Unknown** would indicate neutrality since the author will only be reproducing the words of a third party to inform and not to influence the reader, while any other value would indicate a lack of it.

Title style and stance. The titles of newspaper articles often provide important clues to the reliability of the content. For example, alarmist, subjective or striking titles are suspected of introducing unreliable information. Also, misleading or opaque titles on a topic may indicate clickbait (Zhang et al., 2018). Even certain morphosyntactic features such as the excessive length of a title, the use of more capitalised words (Horne and Adali, 2017) and punctuation marks (especially exclamation marks) and ellipses can lead to a lack of neutrality (Mottola, 2020). In our annotation proposal, these clues are marked in the **TITLE** by means of the attribute-value **style=Subjective**.

Moreover, the stance of the title regarding the news content indicates misleading information when they disagree (Ferreira and Vlachos, 2016). In this case, the existence of the attribute value **style=Disagree** associated with the TITLE label would clearly indicate that the information is Unreliable.

4 Annotation environment and 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, randomly selected and then sorted (36,659 words in total), of which 51 are Reliable and 29 Unreliable, collected from several digital newspapers. Both the reliability of the internal elements and the global reliability of the news item are annotated. News has been annotated with Brat, an intuitive web-based annotation tool (Stenetorp et al., 2012). An example of the graphical annotation in Brat can be observed in Figure 2[3].

Tables 1 and 2 show the total number of labels in the dataset[2].

Label	% Reliable	% Unreliable	Total
WHAT	74.64	25.09	1100
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

Table 1: Dataset description (5W1H 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

Table 2: Dataset description (Structure and EoI labels).

The methodology for creating the dataset followed five steps. First, the dataset was defined and delimited on the basis of three main criteria: domain (health and COVID-19), language (Spanish) and traditional news content structure. Second, news was collected both manually and by means of a web crawler. Third, RUN-AS annotation scheme was applied, and a reliability rating was assigned for each 5W1H label. Fourth, the global reliability of each news item was assigned by two non-expert annotators with knowledge of NLP, taking into account only the plain text, without the labels of the expert annotator. Finally, two inter-annotator agreements were measured to validate the quality of the annotation.

[3] `https://bit.ly/38AyW7K`

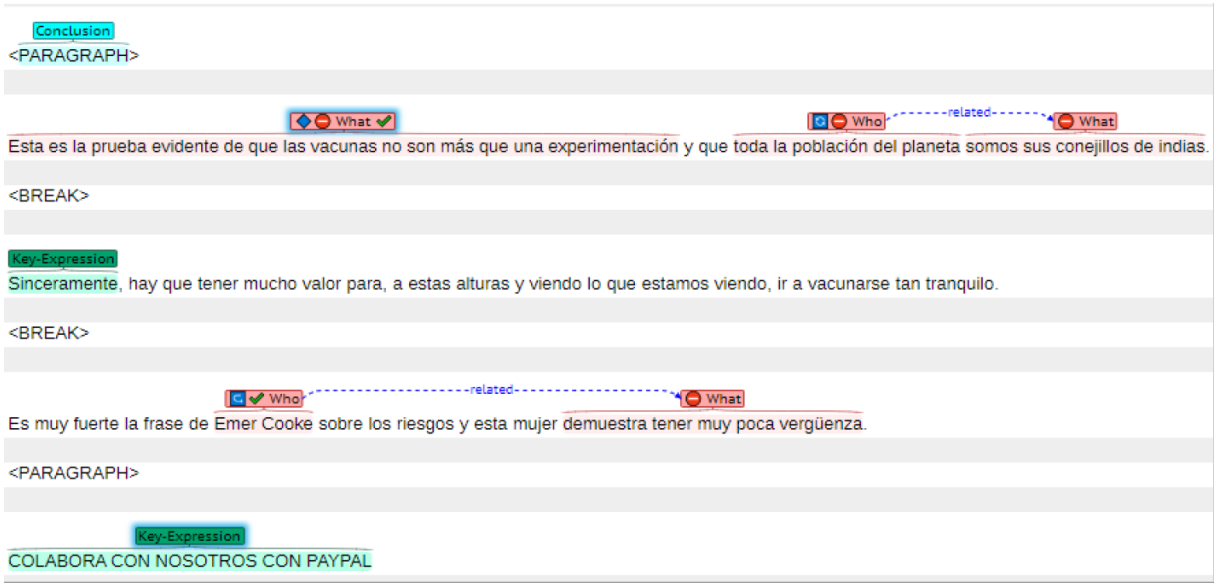

Figure 2: Annotation of 5W1H, Inverted Pyramid and Elements of Interest on Brat.

4.1 Annotation quality

Two inter-annotator agreements were calculated independently using the Cohen's Kappa metric (Vieira, Kaymak, and Sousa, 2010). Firstly, the inter-annotator agreement regarding the three levels of RUN-AS annotation (Structure, 5W1H and EoI) was performed. Secondly, the inter-annotator agreement regarding the annotation of the global reliability of the news item was obtained.

4.1.1 Labels inter-annotator agreement

To measure the agreement in the annotation of the three level labels, the annotation of a set of news items comprising 1,337 words was asked to two non-expert annotators. Without previous training, they had to annotate news according to the annotation scheme proposed. The agreement obtained a score of k=0.80 in the Inverted Pyramid and of k=0.53 in the 5W1H. This inter-annotator agreement allowed us to reach the conclusion that annotating semantic elements has a higher level of difficulty and therefore more intense training needs to be provided to annotators for this purpose.

4.1.2 Global reliability inter-annotator agreement

In order to measure the agreement when annotating global reliability of a news item, two non-expert annotators were used. Their annotations had to be made using plain text only, without labels, and following the reliability criteria defined in the scheme. The agreement obtained in this task was k=0.75 which is considered a fairly high score. When there was no agreement among the annotators, a consensus process was carried out.

5 Validation of RUN-AS scheme: Evaluation framework

Several experiments were conducted to validate our RUN-AS scheme and to support the hypothesis that a fine-grained reliability assessment of the elements in a news story can provide an accurate estimation of its global reliability.

SOTA Machine Learning (ML) and Deep Learning (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. From this fine-grained annotation proposal (Structure, Content, and EoI) two types of features were extracted: numerical and categorical. In total, 42 different features were extracted per news item.

From the Structure level, a total of 7 features were extracted as follows: 5 categorical features that indicate the presence of the news structure parts (TITLE, SUBTITLE, LEAD, BODY and CONCLUSION); and, 2 other categorical features extracted from the attributes of the TITLE (stance and style). Concerning the 5W1H content and EoI 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. 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), and 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 some numerical and categorical features extracted from the TITLE and LEAD of a news piece is presented next.

```
{
    TITLE_style: Objective,
    TITLE_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.

5.1 Experiments

To confirm the suitability of the RUN-AS proposal, we decided to test classic ML algorithms that obtained good results using numerical and categorical features. In addition, a DL language model, which obtained state-of-the-art results in many tasks within NLP, was used to compare the results. The following experiments were carried out:

ML performance: the following ML classification algorithms are used: Support Vector Machines (SVM); Random Forest (RF); Logistic Regression (LR); Decision Tree (DT); 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* library[4]. It can be replicated at the Colab[5] notebook.

DL performance (pre-trained transformer model): the Beto[6] language model based on transformer architecture (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. The architecture of classification presented in Sepúlveda-Torres et al. (2021) is 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 proposed 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*[7] 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[8]. The cross-

[4] `https://scikit-learn.org/stable/`
[5] `https://bit.ly/37KNHnM`
[6] `https://github.com/dccuchile/beto`
[7] `https://simpletransformers.ai/`
[8] `https://bit.ly/3L5LvJg`

validation strategy was performed in all experiments enabling all available data to be used for training and testing (Bergmeir and Benítez, 2012). 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. 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 and a discussion of those results. Table 3 presents the performance of experiments explained in Section 5. All the models that used RUN-AS features 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), and BETO using RUN-AS annotation, obtaining a 0.854 of F_1m. It is noteworthy that when using the whole document annotated with a single reliability value (baselines) the best F_1m value is obtained by AdaBoost with 0.748 F_1m, followed by Random Forest and Decision Tree. However, for the rest of the approaches, the results using the document with a single reliability value are very poor. All approaches are significantly improved by using the information provided by the annotation labels of the RUN-AS scheme. Therefore, these results validate the main hypothesis presented in this research, i.e., that individual 5W1H components reliability are a better predictor of overall news story reliability.

7 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 and its essential parts 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, inter-annotator agreements were measured, both those related to the three-level RUN-AS label annotation as well as those related to the global reliability of the news item. The results indicate the intrinsic complexity derived from a semantically rich annotation scheme. Experiments conducted have shown that the individual reliability of each of the elements annotated contributes to assessing the overall reliability of a news item with a 0.948 F_1m performance. Therefore, the experiments presented here support 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.

This annotation is complementary to other lines of research, such as fact-checking or contradiction detection, as it provides useful information at a first level of a text-only annotation. Our proposal is designed to annotate the style, the structure of the story, the tone, the evidence, the neutrality or the way in which information is provided. These are key characteristics that distinguish Reliable from Unreliable news. As future work, we are developing an assisted annotation methodology that combines both manual and automatic approaches. This semi-automatic system will reduce the time and the effort spent on compilation and annotation tasks, enabling a RUN dataset extension. Furthermore, performance in the veracity detection task of the RUN-AS annotated dataset will be evaluated to determine to what extent reliability detection can support veracity detection.

Acknowledgments

This research work is funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR through the projects “TRIVIAL” (PID2021-122263OB-C22) and “SocialTrust” (PDC2022-133146-C22). It is also supported by Generalitat Valenciana through the project “NL4DISMIS” (CIPROM/2021/21) and Consellería de Innovación, Universidades, Ciencia y Sociedad Digital (ACIF/2020/177).

References

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

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

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	0.650	0.392	0.912	0.875
Decision Tree	0.737	0.683	**0.950**	**0.948**
MLP	0.712	0.570	0.925	0.912
AdaBoost	**0.787**	**0.748**	**0.950**	0.945
GaussianNB	0.612	0.456	0.687	0.570
	Baseline model		**Model with RUN-AS features**	
BETO	**0.850**	**0.800**	**0.887**	**0.854**

Table 3: Experiments results using ML and DL methods.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

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

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

DeAngelo, T. I. and N. S. Yegiyan. 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. and A. Vlachos. 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*, pages 1163–1168, San Diego, California, June. Association for Computational Linguistics.

Figueira, Á. and L. Oliveira. 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., B. D. Horne, and S. Adali. 2018. An exploration of unreliable news classification in brazil and the us. *arXiv preprint arXiv:1806.02875*.

Hamborg, F., C. Breitinger, M. Schubotz, S. Lachnit, and B. Gipp. 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*, pages 339–340.

Horne, B. and S. Adali. 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*, volume 11, pages 759–766.

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)*, pages 1–5. IEEE.

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.

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

Paka, W. S., R. Bansal, A. Kaushik, S. Sengupta, and T. Chakraborty. 2021. Crossstitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393.

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

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

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

Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 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*, pages 2931–2937.

Saquete, E., D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, 141:112943.

Sepúlveda-Torres, R., E. Saquete Boró, et al. 2021. Gplsi team at checkthat! 2021: Fine-tuning beto and roberta. *CEUR*.

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

Shao, C., G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96:104.

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

Silva, R. M., R. L. Santos, T. A. Almeida, and T. A. Pardo. 2020. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 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*, pages 102–107.

Thomson, E. A., P. R. White, and P. Kitley. 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., U. Kaymak, and J. M. Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems*, pages 1–8. IEEE.

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

Vosoughi, S., D. Roy, and S. Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

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

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhang, A. X., A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.

Zhou, X. and R. Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.