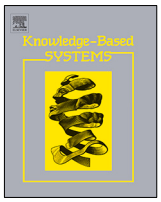
Knowledge-Based Systems 275 (2023) 110723

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



A semi-automatic annotation methodology that combines Summarization and Human-In-The-Loop to create disinformation detection resources

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete *, Patricio Martínez-Barco

Department of Software and Computing Systems, University of Alicante, Spain

ARTICLE INFO

Article history:
Received 2 April 2023
Received in revised form 25 May 2023
Accepted 9 June 2023
Available online 16 June 2023

Keywords:
Natural Language Processing
Semi-automatic annotation
Disinformation detection
Summarization
Dataset construction
Human-in-the-loop Artificial Intelligence

ABSTRACT

Early detection of disinformation is one of the most challenging big-scale problems facing present day society. This is why the application of technologies such as Artificial Intelligence and Natural Language Processing is necessary. The vast majority of Artificial Intelligence approaches require annotated data, and generating these resources is very expensive. This proposal aims to improve the efficiency of the annotation process with a two-level semi-automatic annotation methodology. The first level extracts relevant information through summarization techniques. The second applies a Human-in-the-Loop strategy whereby the labels are pre-annotated by the machine, corrected by the human and reused by the machine to retrain the automatic annotator. After evaluating the system, the average annotation time per news item is reduced by 50%. In addition, a set of experiments on the semi-automatically annotated dataset that is generated are performed so as to demonstrate the effectiveness of the proposal. Although the dataset is annotated in terms of unreliable content, it is applied to the veracity detection task with very promising results (0.95 accuracy in reliability detection and 0.78 in veracity detection).

1. Introduction

In the current digital ecosystem, mis- and dis-information are two highly alarming phenomena, triggering numerous efforts to improve their detection. The main difference between mis- and dis-information lies in the intention [1], because misinformation is not created with the intention to harm but information is eluded or erroneous, whereas disinformation results from a deliberate attempt to deceive or mislead [2]. However, both phenomena result in a common outcome: misleading and confusing information for the audience.

Fake news is a widely disseminated form of disinformation. Fake news is structured and written in a way that makes it difficult to distinguish between what is true or false. Fake information is diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information [3]. Assessing the veracity of information in a news item implies necessarily assessing both content and context signals. Content signals can be determined by only considering the text or content of an article [4,5], whereas context signals require using external world knowledge. Given that this research proposal only uses the news content and no external knowledge is used, rather than focusing on the veracity concept ("the quality of being true"), we deal with the concept of reliability ("the quality of being likely to be correct or true").[1] The difference between both terms lies in the word "likely" included in the definition of reliability. Both concepts are closely related, as fake news often includes a mix of reliable and unreliable information items. Therefore, there is a direct relationship between the veracity of a news item and the reliability of the information contained in it [6]. In view of this connection, our research is grounded in detecting the content reliability/unreliability signals. By determining the reliability of content, readers can be more aware of the information they receive or spread, which may contribute to mitigating the distribution of inaccurate or incorrect information. Reliability can be determined almost precisely by a computer following certain criteria and without the need to have world knowledge that is not always available. However, this does not happen with veracity, which does require external knowledge for a particular news item to be classified as true or false. Reliability would be a prior and supporting step to making a decision on the veracity of a news item because, as this research will demonstrate, reliability detection can assist the task of veracity detection. Verification

* Corresponding author.
E-mail addresses: alba.bonet@dlsi.ua.es (A. Bonet-Jover), rsepulveda@dlsi.ua.es (R. Sepúlveda-Torres), stela@dlsi.ua.es (E. Saquete), patricio@dlsi.ua.es (P. Martínez-Barco).

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

https://doi.org/10.1016/j.knosys.2023.110723

of the facts may take place at a later stage through multiple independent fact-checking organizations [3], which is beyond the scope of this work. Therefore, in this research we focus on detecting reliable and unreliable information in Spanish news texts, without using external knowledge, and based exclusively on the news content and a set of reliability criteria that will be described in the annotation scheme description section.

To deal with unreliable news content from a computational perspective, Artificial Intelligence (AI) and Natural Language Processing (NLP) are required. At present, AI cannot learn by itself and needs to be nourished by examples created by humans [7]. Indeed, when a problem is approached from the AI perspective, either with Machine Learning (ML) or Deep Learning (DL) techniques, millions of instances of human feedback are required to get the training annotated datasets that will be used to train and evaluate the systems that will be in charge of solving the problem [8]. An efficient dataset would be one that can be created as quickly and inexpensively as possible, without deteriorating performance.

Considering this AI context where the need for corpora is essential and highly costly, finding an effective and efficient way to obtain them becomes a fundamental motivation. Furthermore, applying AI to the problem of disinformation, which is increasing in today's society, makes it an important challenge to address.

Therefore, the primary objective of this work is to implement a semi-automatic annotation methodology to support complex semantic annotation of news datasets.

A secondary objective is to fill the gap of the scarcity of training data – especially in languages other than English, such as Spanish – by applying the methodology to create a quality resource capable of modeling the disinformation problem.

To address these objectives, we focus on two important Human Language Technologies. First, automatic summarization is used to reduce the quantity of information requiring annotation to what is relevant. Secondly, the Human-in-the-Loop (HITL) concept is applied to tackle the semi-automatic annotation of the relevant information previously filtered by the automatic summary. The HITL concept is an extensive area of research that covers the intersection of computer science, cognitive science, and psychology. HITL is applied in AI – specifically NLP and ML in this work – because building AI technology with human intervention allows human tasks to be assisted to increase efficiency [7].

In the proposed methodology, summarization and HITL techniques are combined to assist the annotation process, which reduces costs for complex annotations. The combination of these two techniques may reduce the time and expertise required for this procedure.

These objectives make the following novel contributions to the research area:

- A new semi-automatic annotation methodology to minimize the cost of creating efficient and effective complex datasets.
- The application of the methodology to enrich an existing fake news dataset[2] with more complex information following a reliability annotation schema (RUN-AS). The semi-automatic dataset generated[3] is available to the research community.
- An evaluation framework that demonstrates how the resources generated by this methodology contribute to the detection of reliability and fake news in an effective and efficient manner.

[2] https://github.com/jpposadas/FakeNewsCorpusSpanish

[3] Available at https://github.com/rsepulveda911112/RUN-AS-SFN

This paper details the methodology designed to build a semi-automatic annotation tool that comprises two levels. The resulting improvement in annotation efficiency from applying the methodology is measured.

The paper is divided into the following sections: Section 2 presents an overview of the most relevant scientific literature concerning disinformation datasets, automatic summarization and HITL techniques; Section 3 describes the guidelines used in this research; Section 4 presents the methodology proposed; Section 5 describes the dataset construction, following the proposed methodology; Section 6 evaluates the benefits of using the methodology; Section 7 presents the performance of the semi-automatic reliability dataset in reliability detection and veracity detection tasks; and finally, in Section 8, conclusions as well as future work are presented.

2. Background

As our research develops a semi-automatic annotation methodology for a dataset in the disinformation domain, this section deals with the work related to different disinformation datasets and their annotation techniques, automatic summarization, and HITL strategies, the latter being relevant for reducing annotation costs without compromising accuracy.

2.1. Disinformation datasets

In NLP, numerous corpora have been created so that models can learn from real examples of disinformation created by human experts.

After an in-depth study of the different datasets related to the disinformation task, we found that they are mainly datasets that annotate the veracity of a news item as a whole – rather than considering the constituents of a news item, such as the semantic content or the structure – either with a binary veracity value (true/false) [9–14] or by applying a scale with different degrees of veracity as those normally applied in the fact-checking task [15–18]. This single global classification of news with binary or multiple values depends on external knowledge, such as fact-checking platforms [19]. Few datasets use a reliability classification and usually this classification is applied on the basis of the source's credibility [20] and not of purely textual or linguistic characteristics [21].

To the author's knowledge, corpora that address the disinformation task in Spanish are scarce. Those that exist, namely [22, 23], both provide a general overview of the veracity of the whole document, but lack a fine-grained approach to veracity by considering the quality of the essential information within the news piece.

The novelty of our proposal with respect to previous work is twofold. First, is the classification of a news item into Reliable or Unreliable, based on a purely textual analysis without relying on external information. Second, we design a multi-level annotation guideline. This means that a reliability value is assigned to the different parts of the news item, i.e. to the information considered as essential in a traditional news item in addition to assigning one to the whole news item.

2.2. Automatic summarization

Previous research in Text Summarization has been shown to have a positive impact on society since the use of summaries has been beneficial in different areas, such as education – where summaries are used to support reading comprehension tasks [24–27] –, business – by producing, for instance, an automatic summary of event logs to help analysts [28] –, or health, regardless of

whether the summaries were created manually [29,30], or automatically [31]. The benefits of summarization are partly due to their capacity to identify the most relevant information in a document, and condense it into a new text, thereby helping to reduce time and resources when it comes to manage large amounts of data. These methods have proven to be effective when integrated as an intermediate component of more complex systems.

Text summarization approaches can be divided into extractive and abstractive [32,33]. Extractive approaches focus on detecting the most relevant information in a text, which is then copy-pasted verbatim into the final summary [34]. By contrast, abstractive approaches detect the relevant information, and then, a more elaborate process is performed whereby the information is fused, compressed, or even inferred [35]. Besides these two, hybrid approaches have been developed, which combine extractive and abstractive methods [36].

2.3. Human-in-the-loop techniques

According to [37], there is a need for a hybrid model solution that combines the efforts of both humans and machines.

HITL-AI systems continuously improve because of human input, addressing the limitations of previous AI solutions and bridging the gap between machine and human beings [38]. These systems aim to leverage the ability of AI to process huge amounts of data while relying on human intelligence to perform very complex tasks, such as in the case of natural language understanding [39]. The HITL methodology is being used in several studies to increase efficiency in data collection, such as in the cases of [40,41], since "the continuous executive loop develops a more reliable human-AI partnership to a certain extent, contributing to higher accuracy and stronger robustness of the NLP system" [42].

One of the principles of HITL is to assist human tasks with ML to increase efficiency.

A very extended HITL strategy is Active Learning (AL). AL is used when obtaining labeled data demands a large amount of time or money, as AL aims to select examples with high utility for the model [43] and increases the performance of the learning model while reducing the amount of annotated data required [44].

Besides AL, the HITL strategies include two distinct goals that are normally combined: improving the accuracy of the ML application via human input; and, facilitating the human task with the aid of ML. In this work, the latter is our main objective, but the former (the accuracy of the model) will improve as larger datasets are obtained.

HITL-ML has been successfully applied in a variety of areas such as, governmental [45], medical [46], and energy [47]. More specifically, as for applying HITL to dis- and mis-information detection, some works are key. [39] presented the challenges and opportunities of combining automatic and manual fact-checking approaches to misinformation, developing a human-AI framework. This work is more focused on fact-checking and not on reliability. [48] aimed to determine the set of techniques that is best suited for disinformation detection and how each technique might best be used towards this end. However, none of these works leverage the annotation process of a disinformation dataset.

Our work deploys a HITL strategy to enable an increase in the amount of annotated data, thereby reaching the target accuracy more quickly and easily. To thoroughly understand the annotation construction and its complexity on account of its high semantic and linguistic load, the next section presents a brief explanation of the RUN-AS annotation scheme and the reliability criteria adopted.

3. Brief description of RUN-AS annotation scheme

Our work is focused on the annotation of news sourced from a variety of domains and collected from Spanish digital newspapers. Traditional news presents a characteristic structure and some essential elements that follow two well-known journalistic techniques: the Inverted Pyramid and the 5W1H.

Regarding the journalistic structure, well-built news tends to present five common parts which are the TITLE, SUBTITLE, LEAD, BODY and CONCLUSION. In addition, these parts are placed in order of relevance following a journalistic structure known as the Inverted Pyramid. This structure is characterized by placing the most important information at the beginning of the news article, while the least relevant information is located at the end [49].

In terms of content, well-built articles present semantic information through a journalistic concept known as the 5W1H: WHO?, WHAT?, WHEN?, WHERE?, WHY?, and HOW?. These questions allow the extraction of semantic information related to a news item and "are essential for people to understand the whole story" [50]. An example of annotation is presented in Fig. 1.

Using these two concepts, a fine-grained annotation guideline called RUN-AS (Reliable and Unreliable News Annotation Scheme)[4] has been designed. The novelty of this annotation scheme lies in the reliability classification based on a purely textual, linguistic and semantic analysis (without depending on external knowledge). RUN-AS presents three levels of annotation: structure (Inverted Pyramid), content (5W1H) and Elements of Interest (textual clues about formatting or phraseology that enables the detection of suspicious information). Furthermore, each essential content (5W1H) item is annotated by assigning them an individual reliability value and thus differentiating between reliable and unreliable information included in the same news item. The reliability criteria presented in this proposal focus on the principles of accuracy and neutrality:

- **Accuracy:** One of the key factors in determining the reliability of information is accuracy. 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 generalize or to provide inaccurate data. For example, a reliable WHEN is: "el viernes 19 de marzo" *(on Friday 19 March)* whereas "hace mucho tiempo" *(a long time ago)* lacks of accuracy. On the contrary, the presence of figures indicates accurate information that can be easily fact-checked with external sources, thus denoting reliability, for instance "se han administrado 6.000.000 dosis de vacunas" *(6,000,000 doses of vaccine have been administered)*.
  Lack of information: The absence of important data in the text (such as the cause of or reason for an event, the subject of the action, etc.) or the lack of evidence such as scientific studies or official and verified data denotes unreliable information. Sometimes, the author states that the information is based on scientific studies without specifying which ones, which gives it little credibility. The lack of data and sources is another typical characteristic of disinformation, turning news into stories that lack informative content. For example: "según algunos científicos" *(according to some scientists)*.
  Orthotypography: There is a negative reliability impact when there are spelling mistakes, poor or careless writing style,

[4] https://gplsi.dlsi.ua.es/resources/NewsReliabilityAnnotation

<TITLE><WHAT>Fatal accident</WHAT> in <WHERE>Madrid</WHERE></TITLE>

<LEAD><WHEN>Last night</WHEN> <WHAT>a traffic accident</WHAT> took place in <WHERE>Madrid</WHERE> <HOW>after a car crashed into a shop</HOW>. <WHO>The driver</WHO> <WHAT>lost control</WHAT> <WHY>due to the effects of alcohol.</WHY></LEAD>

<BODY> [...] </BODY

Fig. 1. Example of the 5W1H labels annotation.

inadequate punctuation or constant use of capital letters. For instance, "aquí en nuestro Pays" *(here in our "Countri")* denotes unreliable information.

- **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), considered in the RUN-AS schema, are the following:
  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 empathize [51]. Furthermore, 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 [6]. Some examples 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 quotes add neutrality to a news item since it indicates that the information comes from an external source [6]. However, when the author is clearly in favor or against the quote, an important hint of subjectivity is introduced.
  Title style and stance: The titles of newspaper articles often provide important clues as 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 [6]. Even certain morphosyntactic features such as the excessive length of a title, the use of more capitalized words [52] and punctuation marks (especially exclamation marks) and ellipses can lead to a lack of neutrality [53]. Moreover, the stance of the title regarding the news content indicates misleading information when they disagree.

Our proposal does not present a veracity classification into fake or true information. We provide a reliability classification based on a purely linguistic and semantic analysis that takes into account several elements such as vagueness, subjectivity, lack of evidence or emotionally charged content that influences reader's opinions and feelings. Furthermore, one of the future benefits of this fine-grained annotation proposed by RUN-AS is providing an extra level of explainability for the predictions obtained.

4. Design and implementation of the semi-automatic annotation methodology

For the purpose of simplifying the dataset generation, a novel methodology of semi-automatic annotation is proposed. This methodology combines automatic relevant information extraction through summarization approaches as well as a HITL strategy to automate pre-annotation of the dataset. The final aim is to minimize the effort required by the human participant in the annotation via AI, thereby creating larger and less costly datasets. At the same time, the human is enhancing the AI training as human-corrected annotations are reused to retrain AI models.

The semi-automatic annotation methodology consists of two levels, in which the most relevant information of each news item is selected, by first using summarization techniques and second, a HITL strategy is applied for automatic pre-annotation. The relevant information is pre-annotated with the news structure and the 5W1H, and this pre-annotation is provided to the human annotator, who corrects and completes the annotation. In a loop procedure, this human feedback is used to re-train the pre-annotation model. Fig. 2 shows the methodology and steps followed to obtain the semi-automatic annotated dataset.

4.1. Level 1: Automatic relevant information extraction

Due to the semantic complexity of our annotation scheme, instead of annotating the whole document, we are applying summarization techniques to extract the relevant information of the document (Step 1 in Fig. 2). The summarized news items are stored (Step 2-Fig. 2) and used as input to the next level (Step 3-Fig. 2).

In the implementation of the methodology we decided to use the popular and effective TextRank extractive summarization algorithm [54], given its good performance, execution time and implementation availability.[5] This algorithm represents the input text as a graph, where the vertices are the sentences to be ranked, and the edges are the connections between them. Such connections are determined by the similarity among the text sentences measured with respect to their overlapping content. Then, a weight is computed for each of the graph edges indicating the strength of the connection between the sentences pairs/vertices. Once the graph is built, a weighted graph-based ranking is performed in order to score each of the text sentences. The sentences are then sorted in descending order according to their score. Finally, the top ranked sentences, in our case ten, are selected to be included in the final summary.

4.2. Level 2: Automatic pre-annotation and human-in-the-loop

Once the relevant information is obtained, in this level an automatic pre-annotation of both the structure and 5W1H labels is carried out by the system to assist the expert (Step 4-Fig. 2). The benefit of this is that the annotator does not need to label from scratch, but simply revises and completes the pre-annotation done by the system (Step 5-Fig. 2). The structure of news has been annotated by a rule-based system that was developed following the Inverted Pyramid theory. In the case of 5W1H labels, a DL model previously trained with 5W1H labels examples (pink cylinder, represented in Fig. 2) was used.

[5] https://pypi.org/project/sumy/

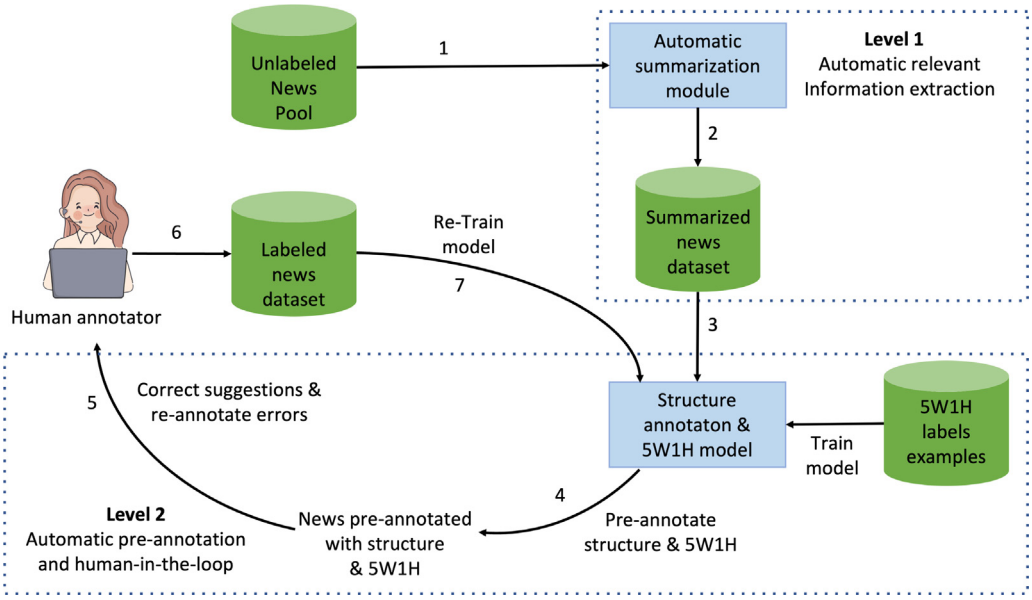


Fig. 2. Design of the semi-automatic annotation methodology.

The human annotator checks that the pre-annotation proposed by the semi-automatic system meets the criteria of the dataset (Step 6-Fig. 2). Furthermore, the pre-annotated items according to the 5W1H model are edited by the annotators, and the rest of RUN-AS annotation is added. Finally, a new annotated batch is added to the dataset (Step 6). This labeled dataset is used not only to train reliability detection models but also to re-train the 5W1H model, thus closing the human-machine loop (Step 7-Fig. 2).

4.2.1. Implementation details of 5W1H model

To deal with this task, initially, a Keyphrase Extraction system was used and a model was trained from scratch to detect the span of the 5W1H. However, this approach did not obtain good results, so it was decided to try Question Answering techniques as they are closely aligned to the task. To perform the automatic pre-annotation of the 5W1H, a question answer **(QA) pre-trained model** available at Hugging Face[6] is used. This model was built with a fine-tuned version of BETO model [55] on SQuAD-es-v2.0 dataset [56] to fit in QA task.

This **QA pre-trained** model was re-trained (twice in our case) to detect 5W1H labels, which is known as fine-tuning. A news dataset with 170 items, previously annotated with 5W1H annotations[4] (RUN Dataset), was used as initial examples of training. This dataset is partitioned into three sets (training, validation, and test). The training and validation sets are updated with the new 5W1H examples annotated in the HITL process. This process allows us to improve the 5W1H model with more annotated news batches. For re-training the model, the following inputs are used: questions (5W1H); question context; and, their respective answers. Next, the model returns an answer as well as a score that represents the probability of certainty associated to the answer.

In this HITL process, firstly, the **M_1** model is obtained after fine-tuning on initial annotated news (RUN dataset). After the first loop, the 5W1H model was re-trained with 250 more news items obtained with the proposed methodology (150 from the training set and 100 from the test set), obtaining the second 5W1H model (**M_2**).

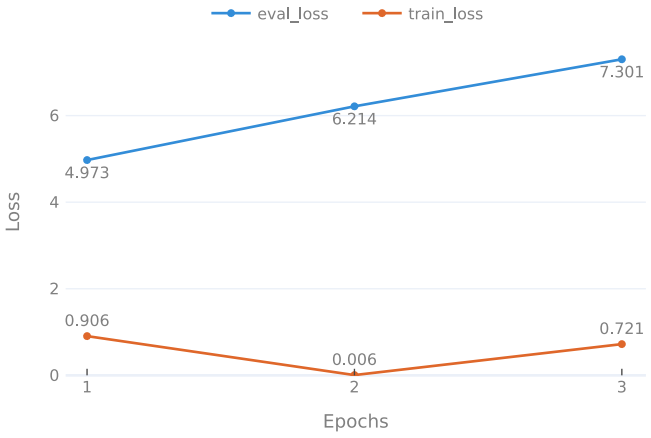


Fig. 3. Loss curve using the training and development sets during training of M_2.

The training process was carried out using the Simple Transformers library.[7] The initial hyperparameter settings for the fine-tuning are: maximum sequence length of 128; batch size of 8; training rate of 4e-5; and, training performed over 3 epochs. This model can be replicated at Github.[8]

Fig. 3 shows the loss curves for training and evaluation where the behavior of the model can be seen during 3 epochs of training.

According to the graph in Fig. 3, after the first training epoch, the loss in the training curve decreases from 0.906 to 0.006, and the loss in the validation curve increases from 4.973 to 6.214 (second epoch). This behavior is constant in the third epoch, which indicates that the model is overfitting after the first epoch. Therefore, we selected the model training with one epoch to pre-annotate the 5W1H labels. Both fine-tuned models have the same hyperparameters and the training and validation curves describe a similar behavior, overfitting after the first epoch.

The QA models obtained in each loop are evaluated and the following five metrics are obtained:

6 https://bit.ly/3zfnisx

7 https://simpletransformers.ai/

8 https://github.com/rsepulveda911112/BETO_QA_SPANISH_5W1H_fine_tuning

Table 1
Comparison between the QA model with and without fine-tuning.

Model	EM	Similar	Incorrect	Overall EM	F_1
QA pre-trained model	30	396	141	5.29	0.19
M_1 fine-tuning	263	152	152	46.38	0.64
M_2 fine-tuning	**272**	**162**	**133**	**47.97**	**0.66**

- Exact Match (EM): the number of the exact matches of the predicted answer with the manual answers. For example, *scientists* annotated as WHO by the QA model which is correct because it is a subject.
- Similar: the number of the partial matches of the predicted answer with the manual answers. For example, *Spain* annotated as WHERE by the QA model. The choice is correct because it denotes a place, but in a certain context, such as *Spain decreed...*, there is a personification and the country functions as the subject of the sentence, so it is a WHO label.
- Incorrect: the number of predicted answers that do not match the manual responses. For example, *two years ago* annotated as WHERE by the QA model which is incorrect because it is a label of time and not of place.
- Overall EM: EM percentage over the number of predicted examples.
- F_1: The F_1 score is the harmonic mean of the precision and recall [57]. The number of shared words between the prediction and the truth is the basis of the F_1 score.[9]

Table 1 shows the performance of the 5W1H models used to predict the test set. The results of the prediction of the **QA pre-trained** model without 5W1H-fine-tuning are also included.

As shown in this table, both QA models with fine-tuning (**M_1** and **M_2**) obtain better results on each metric. We consider that the most important metric in our pre-annotation task is the Overall EM, because maximizing this metric can help manual annotators to reduce, discard, and modify examples. These results confirm that fine-tuning is beneficial for the 5W1H label pre-annotation task, and demonstrates that the system is enhanced in each retraining loop.

The **M_1** shows an outstanding performance versus the **Pre-training** model. The **M_2** attained the best results in terms of EM, Overall EM and F_1 (Table 1). This finding confirms that with a greater number of examples of the 5W1H labels, a model with high precision can be obtained. This reduces annotation times and assists the human annotator in this complex task.

5. Semi-automatic reliability resource

To assess whether identifying the reliability of the different parts of a news item contributes positively to the task of detecting the veracity of a news item, we used a Spanish dataset created for the detection of fake news, namely, The Spanish Fake News Corpus (SFN) [58]. The SFN corpus contains a collection of 971 news items (491 TRUE and 480 FAKE) compiled from several resources on the Web.[10] In this dataset, the annotation consists of a single veracity value for each news item.

A subset of 400 news items was selected for our proposal (around 50% of the dataset). Preserving the initial annotation of the dataset, the semi-automatic annotation applying RUN-AS scheme was performed obtaining the new RUN-AS-SFN dataset (a subset of SFN annotated following RUN-AS annotation scheme).

Table 2
Numerical description of the RUN-AS-SFN annotated dataset.

Sets	Veracity		Reliability	
	True	Fake	Reliable	Unreliable
Training	143	157	160	140
Test	48	52	62	38
Total	191	209	222	178

Table 3
Numerical description of the 5W1H items in RUN-AS-SFN dataset.

5W1H	Unreliable	Reliable	Total
WHAT	1,465	2,670	4,135
WHEN	133	1,200	1,333
WHERE	103	1,543	1,646
WHO	521	3,588	4,109
WHY	324	512	836
HOW	194	568	762
Total	2,740	10,081	12,821

From the subset, 300 news items were used for training and 100 news items were used for testing. The following topics are covered in the RUN-AS-SFN annotated subset: Economy, Sports, Science, Education, Health, Society, Entertainment, Politics and Security. The figures regarding the dataset after the semi-automatic RUN-AS annotation procedure are presented in Table 2.

As explained in the methodology, the annotation of the different parts of the news item was performed only for the relevant information extracted in the summary. However, a global reliable value is given to each news item, and for this annotation the entire news content is considered. As presented in Table 2, the subset selected was balanced in fake and true news items, both for the training and for the test. After performing the annotation of reliability, the number of reliable versus unreliable news items was quite balanced, with only slightly larger reliable items than unreliable ones.

Previous research found that news mixes unreliable and reliable information, which hinders the disinformation detection task [59]. This is why it is important that the different parts and essential content of a news item have specific reliability values, which influence the global reliability value of a news item. Details regarding the reliable and unreliable 5W1H items in the whole RUN-AS-SFN annotated dataset are presented in Table 3.

As shown in Table 3, the number of reliable 5W1H is much higher than that of unreliable, despite the fact that reliable and unreliable news are quite balanced. This is due to the fact that fabricated news is of increasingly better quality, incorporating more and more reliable information, which makes the detection of the veracity of a news item even harder.

5.1. Annotation process

Two experts performed the annotation task. The expert annotators are linguistic researchers specialized in NLP (1 PhD annotator who is the author of the annotation guidelines and 1 Ph.D. student) and both are native Spanish speakers.

The user interface is the Brat tool,[11] which together with the implemented assisted system enables the annotator to discard, accept or modify the annotation labels. Fig. 4 shows the annotation interface. There is a pop-up window for editing a specific label, that in this case shows an example of a WHEN annotation label in the Spanish text *A las 7 en punto en la mañana del 25 de enero* (At 7 o'clock on the morning of January 25). The interface highlights the sentences belonging to the summaries. The

[9] precision is the ratio of the number of shared words to the total number of words in the prediction, and recall is the ratio of the number of shared words to the total number of words in the ground truth [56]

[10] https://github.com/jpposadas/FakeNewsCorpusSpanish

[11] https://brat.nlplab.org/

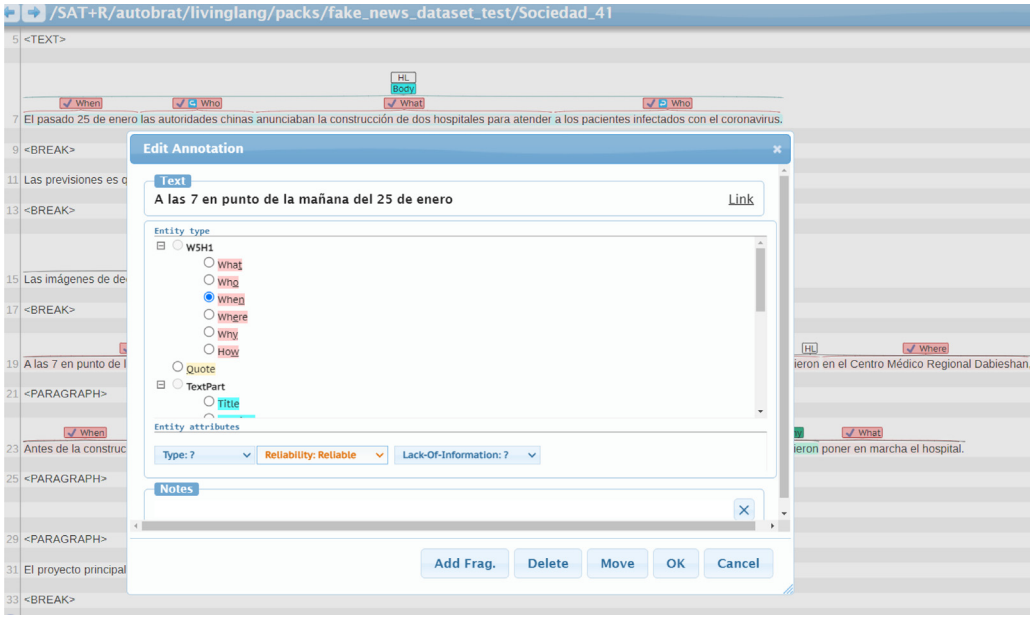


Fig. 4. Table with labels and attributes in Brat.

use of this interface enables the annotator to annotate quickly, accurately and easily. Furthermore, good assisted systems provide annotations that benefit from quality, not just quantity.

The annotation process was performed in 40 sessions of 10 news items annotated per session. The average time used per session was 80 min. After both annotators completed the annotation, its quality was measured as detailed next. Finally, to obtain the final annotated dataset, the annotators compared their annotations, and in the case of disagreement they reached a consensus.

5.2. Annotation quality

The quality of the annotation was measured both quantitatively and qualitatively. Cohen's *kappa* [60] was initially considered to quantitatively assess the quality of the semi-automatic annotated dataset. This renown agreement metric is commonly applied when there are two annotators and it controls for chance annotation agreement. However, Cohen's *kappa* was finally discarded in this research because when tasks require the labeling of boundaries – such as in named entity recognition or the task in this work – there are often a very large number of potential spans that no annotator ever extracts. In this case, the expected chance agreement is effectively zero [61]; and thus, the *kappa* is equivalent to F_1. For this reason F_1 may be more appropriate in these previously mentioned contexts to quantify the inter-annotator agreement (IAA) and therefore, to assess the quality of the annotated dataset [62]. Considering this, our experiment uses F_1 instead of *kappa* to measure the annotation quality.

To obtain F_1-Measure, precision and recall metrics are first calculated. In our case, precision and recall are measured using one annotator as a reference (PhD expert and author of the annotation guidelines) and the other as a prediction. When comparing annotations, for each 5W1H label, there was an agreement between the annotators (A and B) when they agreed to assign the same category (WHO, WHAT, WHERE, WHEN, WHY, HOW) and the same value for the reliability attribute to a text span. For example, *scientists* annotated as WHO unreliable by the annotator A and as WHO unreliable by the annotator B. We also considered for the annotation of text spans partial matches as agreement, due to the semantic complexity of 5W1H labels. Thus, we considered also a match those cases where there was a slight difference in length regarding the span of the elements to be annotated, but the 5W1H label assigned was the same. For example: *scientists* annotated as WHO by the annotator A and *scientists specialized in biophysics* annotated as WHO by the annotator B.

Table 4
IAA based on Precision, Recall and F_1-measure by annotation level of RUN-AS-SFN Dataset.

Annotation level	Precision	Recall	F_1
Inverted Pyramid	0.91	0.62	0.74
5W1H	0.77	0.53	0.60
Elements of Interest	1.00	0.30	0.46
Complete annotation	0.89	0.48	0.60

Given a prediction and a reference, Precision (P) is the proportion of cases that the prediction classified as positive that were positive in the reference. It is equivalent to a positive predictive value.

$$P = \frac{\#TruePositive}{\#TruePositive + \#FalsePositive} \tag{1}$$

Recall (R) is the proportion of positive cases in the reference that were classified as positive by the prediction.

$$R = \frac{\#TruePositive}{\#TruePositive + \#FalseNegative} \tag{2}$$

The two metrics are combined as their harmonic mean, known as the F_1 measure, which is the weighted average of Precision and Recall. Note that inverting the reference and the prediction only inverts the precision and the recall but has no effect on the F_1 measure itself [63]. It can be formulated as follows:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Table 4 presents the Precision, Recall and F1-measure obtained by each annotation level (Structure, essential content and Elements of Interest). Furthermore, an average of all metrics was calculated and presented.

As shown in Table 4, the results indicate a substantial inter-annotator agreement in the structure level, but this agreement decreases in the levels that are semantically more complex such

Table 5
IAA based on Precision, Recall and F_1-measure of the 5W1H labels along with their respective reliability in RUN-AS-SFN dataset.

5W1H	Precision	Recall	F_1
WHAT	0.83	0.73	0.77
WHO	0.84	0.58	0.67
WHEN	0.92	0.72	0.78
WHERE	0.76	0.51	0.58
WHY	0.83	0.53	0.63
HOW	0.42	0.12	0.19
Total	0.77	0.53	0.60

as the 5W1H and the Elements of Interest. Despite the complexity of these two levels, the fact that the 5W1H annotation is assisted helps to improve the agreement in this level. In the case of the Elements of Interest, this annotation is still completely manual and with a high degree of subjectivity which makes it a complex annotation level. In future, semi-automatic annotation at this level will also be considered. As for the 5W1H level, the agreement presented is measuring when there is a match not only in the label but also in the reliability attribute. F_1 was also obtained in the case of only 5W1H label agreement, with a value of 0.72; therefore, the agreement in this case is higher. In addition, since the pre-annotation is performed at 5W1H level, the measures of precision, recall, and F_1 were disaggregated in order to observe the complexity of each 5W1H label separately. The results are shown in the Table 5.

According to the results presented in Table 5, the agreement is considered acceptable for all labels except for the HOW label where it is quite low. Considering these results and in order to detect an inconsistent application of the annotation guidelines or possible reconsideration of the instructions given, it is necessary to identify the patterns and trends in the types of errors made during the annotation process. This is done via a qualitative assessment of the semi-automatic annotated dataset.

At the Inverted Pyramid annotation level, which is related to the structural parts contained in a news item, the most common disagreements between both annotators arises in determining whether the SUBTITLE and CONCLUSION parts should be annotated. Due to the fact that these two parts are optional in the Inverted Pyramid, it is difficult to determine if they are present independently or if the information is included as part of the LEAD or BODY of the news item which are mandatory sections in a well-structured news item. Anyway, in this case, the disagreement is not a serious problem in the end for the classification of the news, since the information contained in one label or another will be finally considered by the system.

As for the 5W1H labels, annotators substantially agree on both the selection of the label type and the reliability of the labels. In addition, the pre-annotation of these labels greatly reduces the risk of subjectivity when selecting the information to annotate. However, it has been observed that, despite pre-annotation, additional labels are found that are not in the reference annotation. This problem occurs especially in complex labels due to their semantics or possible misinterpretation of the guide. One of the most frequently added additional labels that causes the most divergence in the agreement is the HOW label, as indicated in the quantitative analysis (see Table 5). After an analysis of those additional labels, it is revealed that often the second annotator indicates as HOW adjectives or adverbs, such as: *"They needed to work tirelessly"*, where *tirelessly* is annotated as HOW. Since we are only focused on the essential content of the news item, not all the HOW answers referring to an event would be relevant for our task since we are interested in understanding the practical aspects of a situation or the process in which an event has occurred. When this happens, there is a misinterpretation of the guidelines, and for this reason, the criteria for considering specific information as relevant should be more specifically defined. Other elements of disagreement have also been observed with nested expressions, such as *"The Prime Minister of France"*, since they could be considered as a single global label (WHO) or as two distinct labels (WHO and WHERE). For these misinterpretation cases, it is necessary to clearly define in the guidelines that they will be considered a single entity since the first entity is the one that determines the type of the label. Finally, with respect to the error in selecting the type of label 5W1H, discrepancies could be seen in ambiguous cases where both labels could be used, for example: *"In 48 hours"*, which could be interpreted as a HOW or a WHEN depending on the context. As in the previous cases, the analysis of coincidences and divergences in the annotation allows us to consider errors and possible ambiguities and thus define criteria that enable us to refine the annotation.

Regarding the labels that comprise Elements of Interest, given that this level of annotation has not yet been automated, there are discrepancies in the selection of the textual elements to be annotated, as the Recall results indicate. These divergences in annotation are not due to an error in annotating the type of label, since the annotators select the same type of Elements of Interest label whenever they coincide in labeling. However, in this case, the difference is due to subjectivity in choosing which information to annotate and therefore adding more or fewer labels. This leads us to reconsider the instructions given for these types of labels, making them more precise and concrete.

Table 6
Average annotation time per news item by annotation procedure.

Method	Time (min)	Words
Manual Annotation	16.71	510.57
Semi-automatic full news	12.32	482.22
Semi-automatic summarized news	8.07	383.11

6. Measuring improvement in the semi-automatic annotation procedure

To assess the improvement in the annotation process when using the semi-automatic annotation methodology, both the reduction in annotation time and the error rate in the pre-annotation of the 5W1H were measured.

6.1. Measuring time reduction

To demonstrate the efficiency of applying the semi-automatic methodology in the annotation process, we compared the annotation time of the following different options available: (a) manual annotation, (b) semi-automatic annotation with full news items, and (c) semi-automatic annotation with summarized news items. To measure the time-consumption, batches of 10 news items were created for the three previously mentioned annotation options. Average annotation times for each annotation procedure is shown in Table 6.

Considering a similar average word length, ranging from 350 to 550 words, a reduction in the average annotation time per news item can be observed, from 16.71 min/news item for manual annotation, to a reduction to 12.32 min/news item due to the semi-automation of the annotation for the complete news item. The use of summaries for text annotation further reduced the annotation time to 8.07 min/news item. The procedure achieves a time reduction of 50% from a fully manual annotation to this semi-automatic annotation process.

When analyzing news items, there are several factors that influence their annotation, such as the topic. Hence, in addition to

Table 7
Average annotation time (semi-automatic + summarized) per news item by topic.

Domain	Time (min.)	Words
Economy	9.7	444.00
Sports	8.5	326.70
Science	8.75	401.87
Health	9	431.83
Society	7.2	365.80
Entertainment	6	298.60
Politics	7	473.60
Security	7	314.50
Education	5	285.33

Table 8
Ratio of EM, Similar and Incorrect 5W1H labels of the total annotated labels for each M_1 and M_2 models.

5W1H labels	EM		Similar		Incorrect	
	M_1	M_2	M_1	M_2	M_1	M_2
WHAT	0.87	**0.89**	**0.13**	0.09	**0.0**	0.02
WHEN	**0.84**	0.81	**0.04**	0.04	**0.12**	0.15
WHERE	0.65	**0.76**	**0.08**	0.08	0.26	**0.16**
WHO	0.88	**0.96**	**0.07**	0.03	0.06	**0.01**
WHY	0.32	**0.39**	**0.03**	0.03	0.65	**0.58**
HOW	0.43	**0.49**	**0.11**	0.08	0.46	**0.43**

the overall comparison of the annotation time between the three annotation procedures, a more detailed analysis of the selected news items according to topic was carried out. The specific data on the average annotation time per topic are presented in Table 7. Due to their complexity, the topics that made the annotation task most difficult were: Economy (9.7 min/news), Health (9 min/news), Science (8.75 min/news) and Politics (7 min/news). This is because these topics contain numerical data and specific terminology, as well as a denser writing style to present the information more objectively. On the contrary, the topics of Society (7.2 min/news item), Entertainment (6 min/news item) and Education (5 min/news item) are annotated at a different pace because they tend to present information in a more informal style, in addition to the fact that the news items do not require as much prior or specialized knowledge from the reader.

To annotate the reliability of the entire news item requires a reading of the whole item, which influences the annotation time.

It can be observed that the topics which contain more words per news item require more annotation time than shorter ones, i.e., Economy (444 words/news item), Science (401.87 words/news item), Health (431.83 words/ news item) and Politics (473.60 words/news item) versus Entertainment (298.60 words/news item).

Sports news tends to include a lot of quotes and the thread is more difficult to follow than those news items in which the information is presented in a more specific order. In the case of Entertainment, news seeks to distract the reader and inform in a simpler way. Another influential factor in the analysis is the language of the corpus. Most of the news items in the corpus chosen for annotation are written in Latin American Spanish, which may cause some comprehension difficulties for the annotator (with a Spanish linguistic profile from Spain) and delay the annotation of certain topics with cultural information related to sportsmen, politicians, celebrities or society. However, although this may limit comprehension and slow down the annotation task, it also allows the annotator, having less knowledge of that culture, not to be influenced by the context or already acquired knowledge of the world, which allows him/her to be more objective when classifying the reliability of the data.

6.2. Measuring pre-annotation error rate

An analysis of the errors committed by the pre-annotation models of the 5W1H was done to ascertain which labels should be automatically pre-annotated or not. If the annotator has to correct a very high percentage of a type of label, it may be more convenient for this not to be pre-annotated, because correcting a type of label that fails may always be more time consuming than annotating it from scratch.

Table 8 shows the ratio taking into account *Exact, Similar and Incorrect Match* with the total of 5W1H labels pre-annotated with the **M_1** model and the **M_2** model after a loop retrain. Furthermore, two measurements of the pre-annotation ratio were done. To perform this measurement, eight news items were selected and pre-annotated with **M_1**, counting the three categories mentioned above. After annotating 150 news items for training and 100 news items for test set, they were used to re-train the 5W1H model without the eight news items selected for the measurements. Then, the **M_2** model was obtained, and the second loop was started. Finally, the second measurement was performed using **M_2** model for the same pre-annotation.

According to the results presented in Table 8, using M_2, the pre-annotation precision improves significantly, as indicated by the manual recount of labels marked as *EM*, *Similar*, and *Incorrect*. For those marked as *EM*, all labels improve except the WHEN label (which decreases by 0.03 points), with 0.1 being the greatest increase in points in the case of the WHERE label. For the labels marked as *Similar* there was no improvement, which is not significant as it is probably due to a lot of examples being marked as *Similar* after the M_1 prediction model changed to *EM* in the M_2 prediction model. Finally, the ratio of the labels marked as *Incorrect* in the M_2 prediction model decreased in the case of WHERE, WHO, WHY, and HOW labels. Although there is a small increase in the ratio of the WHAT and WHEN labels, the fact that the system is able to detect significantly more labels correctly (more *EM*) carries more importance since it is easier to remove an erroneous label from the annotation than to have to do a complete annotation from scratch. Furthermore, as presented in Table 1, the overall *EM* and F_1 of the M_2 model surpassed the M_1 model.

Finally, after this analysis, we can conclude that the QA model assists the annotator properly, and it is feasible to pre-annotate all 5W1H labels this way. However, we would improve the QA model in the future to reduce the error-rate.

7. Evaluation

In order to evaluate the effectiveness and efficiency of the methodology presented as well as the quality of the semi-automatic dataset generated for the disinformation task, two evaluations are proposed. First, an evaluation of whether the features annotated by applying the RUN-AS annotation enable the reliability/unreliability of a news item to be determined. Hereafter, the reliability annotation will be used to determine whether it is useful for ascertaining the veracity of a news item, when applied to the task of fake news detection. Finally, an analysis of the relationship between veracity and reliability is performed.

7.1. Performance results for the reliability detection task

Several experiments were conducted to validate the semi-automatic methodology proposed. This also supports the claim that a fine-grained reliability assessment of the elements in a news story can provide an accurate estimation of its global reliability.

Table 9
Experiment results using classical ML and DL approaches for Reliability detection task.

Experiments	Baseline NO features		Baseline with RUN-AS features	
	Acc	F_1m	Acc	F_1m
MLP	0.49	0.49	0.93	0.92
SVM	0.62	0.38	**0.95**	**0.95**
LR	0.38	0.37	0.94	0.94
DT	0.41	0.40	0.87	0.86
RF	0.52	0.52	0.88	0.87
RoBERTa	0.71	0.67	**0.77**	**0.74**

According to the literature for disinformation detection, both traditional ML [64,65] and DL [66] based models are used. Considering this, to conduct the experiments, two baselines are proposed. The first baseline uses classical ML algorithms, widely applied in the disinformation classification task. This was used to determine if the semi-automatic annotated dataset is feasible to address the content reliability detection task. The ML approaches used are: Support Vector Machines (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Multilayer perceptron (MLP). This baseline has as input the news content (TITLE and BODY text) encoding in TF-IDF type vectors. We chose this encoding vector because, although it is a classic and essential word representation, it still performs well in the fake news detection task when used with traditional ML algorithms [67,68]. The second baseline is based on DL, using the Transformer architecture to encode the news content (TITLE and BODY text). For this baseline, we used a pre-trained Spanish version[12] of RoBERTa model,[13] followed by a neural network to perform the classification.

In order to evaluate the proposed baselines, a validation set was created from the training set (300 news items) using 20% of the examples. The baseline models are trained and validated with the RUN-AS-SFN annotated dataset to predict the 100 news items that comprise the test set. Each baseline can also be tested using the features obtained from the RUN-AS annotation. For the first baseline, the TF-IDF vectors are concatenated with 42 numerical and categorical features extracted from the RUN-AS annotation. In the second baseline, the texts encoded with the RoBERTa model are concatenated with the external features and passed to the classification neural network—in this case, a multilayer perceptron (MLP).

The features are extracted from the three annotation levels (Structure, Content, and Elements of Interest). 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 were 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), 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 the numerical and categorical features extracted from the TITLE and LEAD of a news piece is presented next.[14]

```
{
  TITLE_style: Subjective,
  TITLE_stance: Disagree,
  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: 0,
  LEAD_WHAT_Unreliable: 2,
  LEAD_WHO_Reliable: 1,
  LEAD_WHO_Unreliable: 0,
  LEAD_WHERE_Reliable: 0,
  LEAD_WHERE_Unreliable: 3,
  LEAD_WHEN_Reliable: 2,
  LEAD_WHEN_Unreliable: 0,
  # ...
}
```

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 appears 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.

Table 9 shows the results of the baselines in terms of the metric F_1m and accuracy (Acc) to predict the test set. In order to evaluate the importance of the annotation used, each baseline is trained only with text (Baseline NO features) and using RUN-AS features concatenated with the encoded vectors of the text (Baseline with RUN-AS features). To replicate the results of each baseline you can use the following GitHub repositories (ML-based baseline[3] and DL-based baseline.[15])

As can be observed from the results presented in Table 9, the best results are obtained with SVM approach (0.95 accuracy and F_1). This evaluation shows that for all ML and DL approaches used, a very high percentage increase in reliability detection results is achieved when the model uses RUN-AS features, despite the fact that the annotation of these features is being performed only on the essential news content provided by the summary.

For baseline systems that do not use external features, the RoBERTa model obtains the best classification results, confirming the power of the pre-trained models when using text only versus the classical ML algorithms. When the features are concatenated with the RoBERTa encoded vectors, the second baseline improves the results, evidencing also the influence of the external features. Despite this result, they are not better than those achieved by classical ML algorithms using features.

Therefore, it can be concluded that both the methodology using the summary and the generated semi-automatic dataset are highly effective for the reliability detection task.

12 Available at https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

13 For more detail about the RoBERTa model, consult [69,70]

14 Only some of the features are shown to exemplify the generation of these features

15 https://gplsi.dlsi.ua.es/resources/BETO_RUN_AS

Table 10
Reduction percentage of each type of 5W1H label with the application of abstracts.

5W1H	Reduction %
WHAT	48.18
WHO	39.65
WHEN	36.84
WHERE	34.61
WHY	53.84
HOW	43.33
Total	42.63

Table 11
Experiment results using classical ML and DL approaches for fake news detection task.

Experiments	Baseline NO features		Baseline with RUN-AS features	
	Acc	F_1m	Acc	F_1m
MLP	0.56	0.44	0.74	0.74
SVM	0.52	0.34	0.75	0.75
LR	0.52	0.44	0.77	0.77
DT	0.54	0.52	0.72	0.72
RF	0.60	0.58	**0.78**	**0.78**
RoBERTa	0.60	0.57	**0.70**	**0.70**

7.1.1. Analysis of error propagation during annotation process

Since the annotation is semi-automatic, the possible propagation of errors produced by the automation tasks will need to be analyzed by each level.

- *Impact in performance derived from Level 1: Automatic relevant information extraction*
  Firstly, the impact of reducing information when applying summarization is analyzed. In order to determine the reduction percentage of annotation in comparison with complete news item annotation, we measured the annotation of a batch of 10 random news items that were fully annotated and the same batch annotated following our semi-automatic methodology. Table 10 presents the average percentage of reduction of 5W1H labels annotated for each type when summaries are applied to the document.
  According to the table, there is an average reduction of 42.63% of the labels compared with a complete annotation of the document. This implies that the training information is reduced by almost half. However, the results in performance presented in Table 9, despite the reduction, are in line with those obtained for this type of task when using the same annotation schema but annotating the complete news item (F_1 0.95 for best ML approach) [71]. This indicates that the choice of using summaries is a feasible solution, since by determining the relevant information, the system learns effectively even if the number of labels per news item is highly reduced.
- *Impact in performance derived from Level 2: Automatic pre-annotation.*
  Considering the errors of the pre-annotation presented in Table 8, the most problematic labels are WHY and HOW, in which the percentage of success and error is the same. This would mean that half of the WHY and HOW labels will have to be modified or removed by the human annotator. In our case, according to the RUN-AS-SFN dataset figures, the percentage of these labels is much smaller than for the other types of labels, being 6.5% of WHY labels and 5.9% of HOW labels in the dataset. This indicates that although it would be necessary to improve the results of the pre-annotation system for these labels, it is currently a minor problem. Furthermore, due to the HITL based methodology, the incorrect or similar annotations are revised and corrected by the human annotator, preventing error propagation issues in the training process.

7.2. Performance results for fake news detection task

To determine how the reliability annotation supports the veracity detection of a news item, we apply the RUN-AS-SFN annotated dataset to the task of fake news detection. The results will be compared with those of the state of the art in the literature on The Spanish Fake News Corpus, which is used as the basis of this research. We performed the same experiments explained in the previous section but in this case to predict the veracity of the news. The results are presented in Table 11.

Table 11 indicates the best accuracy and F_1m results are obtained with RF (0.78 and 0.78 respectively). Using the RUN_AS features implies a clear improvement in the task, with an increase of 0.2 points in F_1m for the RF approach and 0.33 points in F_1m for LR approach. Similar to the experiments in the previous section, the baseline that used RoBERTa model gets very competitive results without features, but there is a lot of room for improvement when external features are concatenated.

Regarding the comparison with the state of the art, one approach trained a classifier to generate a model that can distinguish between real and fake news and experimented with four ML classifiers: SVM with linear kernel, LR, RF, and boosting (BO) [58] . They used two feature representations: one of them is the standard bag-of-words (BOW) model and the other two representations are the character n-grams and POS tags n-grams representation. This previously cited work presents the accuracy obtained in the test set when they trained the classifiers on individual feature sets, such as BOW and POS, and combined those feature sets. Their best result was 0.77 accuracy with RF and BOW+POS. Our experimentation surpasses this result because we obtain 0.78 accuracy with RF when applying the reliability annotated features, and it is important to emphasize that only essential content is annotated and not the entire news body text. These results validate the claim that the annotation of news reliability supports the task of detecting fake news and disinformation.

Since the results reported by [58] are performed over the entire Spanish Fake News Corpus, and considering that we only used a subset of it (50% of the news items), for a fair comparison we replicated the LR model with BOW, in the same way as described in [58]. In this case, the training set is 80% of 300 news items annotated (240 news items), with 60 news items to validate, obtaining an accuracy of 0.68. In our experimentation, using RUN-AS features, LR obtains 0.77 accuracy. However, it is important to emphasize that the replicated accuracy is lower than the one reported by [58] due to the fact that in our replication we used less than half of the news items from the training set and the specific configuration of hyperparameters used by [58] was not reported.

7.2.1. Reliability-veracity relationship analysis

There is a general consensus that fake news contains both reliable and unreliable information and there are linguistic patterns that can add or detract from the reliability of the news. Therefore, we have applied our annotation scheme – which classifies news as reliable and unreliable according to the reliability of its parts – to a published corpus that only provides a global classification of true and false.

In this work we analyze the reliability-veracity relationship via matches and divergences in the RUN-AS-SFN annotated dataset.

Table 12
Relation between Reliability-Veracity in the training set.

Training	True	Fake
Unreliable	15	125
Reliable	128	32

Table 13
Relation between Reliability-Veracity in the test set.

Test	True	Fake
Unreliable	3	35
Reliable	45	17

We consider that there is an annotation match when a news item is reliable-true or alternatively unreliable-fake. We consider that there is an annotation divergence when a news item is reliable-fake or alternatively unreliable-true.

As shown in Table 12, in the training annotated set, 253 news items match, while 47 items diverge. In Table 13, in the test set, 80 news items match whereas only 20 news items diverge.

The news items that diverge between our reliability and the original dataset veracity classification were analyzed in detail to find out why this divergence occurs and some examples to illustrate the situation are presented. Both the training and the test news items with divergence in annotation were thoroughly analyzed.

Firstly, the divergence **Unreliable-True (U-T)** was studied. This means considering a news item as Unreliable when classified as True by Posadas' annotation. In line with the RUN-AS scheme, the following real examples that were originally annotated as True, were subsequently annotated as Unreliable according to the reliability criteria:

- Presence of titles that are not very objective, poorly constructed, unfinished or of the clickbait type, created to attract the user's attention:
  - Example: Ponen en duda investigación de Carlos Trejo sobre "Cañitas" **!y cuentan la verdad!** (*Carlos Trejo's research on "Cañitas" is questioned* ***and they tell the truth!***)
  - Example: **lo que NBC no mostró** de la entrevista con Putin (***what NBC did not show*** *from Putin's interview*)
- Imprecision or vagueness of information, such as impersonal structures, lack of clarity of the subject or imprecise moment in time:
  - Example: **Hace unos meses** (*A few months ago*)
  - Example: **Varios años más tarde** (*Several years later*)
  - Example: Incluso **se ha comentado** que la luchadora no será sancionada [...] (***It has even been commented*** *that the fighter will not be sanctioned [...].*)
- Content that influences the annotator's neutrality, because linguistically the annotated information is objective and well presented, but the content itself seems not very credible or the annotator has knowledge of the world that does not permit objectivity and influences the annotation:
  - Example: **Niño de 8 años se prepara para entrar a la universidad** (*8-year-old boy prepares to enter college*)
- Exaggerated information, for example, with the use of superlatives:
  - Example: [...] retrataba **el caso más escalofriante ocurrido** en una casa en la Ciudad de México (*[...] portrayed* ***the most chilling case that occurred*** *in a house in Mexico City.*)
- Personal remarks through the use of the first person or personal experiences:
  - Example: La secretaria **me dijo** que se trataba de una ocasión [...] (*The secretary* ***told me*** *that it was an occasion [...].*)
- Presence of key expressions that try to influence reader opinion or incite them to spread and believe the information:
  - Example: **Comparte este contenido** (*Share this content*)
- Subjectivity shown from information addressed to the reader, such as questions, personal opinions or unscientific advice and recommendations.
  - Example: **Si estás embarazada**, el mejor remedio ***para olvidarte*** de las náuseas es consumir una pequeña dosis [...] (***If you are pregnant****, the best remedy to get rid of nausea is to consume a small dose of [...]*)
  - Example: **Si no lo crees**, te decimos cuáles son estos beneficios. (***If you do not believe it****, we tell you what these benefits are.*)

  In many cases, those unreliable phrases that bring subjectivity or polarization to the discourse are mixed with reliable and complete information, as seen below:
  - Example: En el periodo comprendido entre el 3 de febrero de 2016 *y* el 30 de mayo de 2019, el Gobierno de España ha concedido, a través del Ministerio de Asuntos Exteriores, Unión Europea *y* Cooperación, 1.884 suvenciones por valor de más de 630 millones de euros (631.179.143, 19 euros). **El chollo de ser feminista en España**. (*In the period between February 3, 2016 and May 30, 2019, the Government of Spain has granted, through the Ministry of Foreign Affairs, European Union and Cooperation, 1,884 grants worth more than 630 million euros (631,179,143, 19).* ***The benefits of being a feminist in Spain***)
- Lack of scientific evidence or sources, which makes the information less credible as we do not know where it comes from or what it is based on:
  - Example: **Diversos estudios** han demostrado que [...]. (***Various studies*** *have shown that [...].*)
- Typographic errors, misuse of capital letters, grammatical errors or even the use of suspension points to generate doubt in the reader.
  - Example: Algunos se pasan de frenada en eso de la reivindicación de derechos e interpretación de la ley **...** (*Some people go too far in claiming rights and interpreting the law* **...**)
  - Example: **AUSENCIA DE INDÍGENAS DEL CAUCA Y CAQUETÁ POR MINGA, PERMITIÓ A ANTINARCÓTICOS DESTRUIR 63 LABORATORIOS DE COCA** (*LACK OF CAUCA AND CAQUETA INDIGENOUS PEOPLE IN CAUCA AND CAQUETA FOR MINGA, ALLOWED ANTINARCOTICS AGENTS TO DESTROY 63 COCA LABORATORIES*)

Secondly, the **divergence Reliable-Fake (R-F)** is studied. This means assessing as a Reliable news item one classified by Posadas' annotation as Fake.

After studying the news items showing this type of divergence, Reliable labels far outnumber Unreliable labels, reflecting the objectivity and neutrality of the information shared applying the RUN-AS scheme reliability criteria. In these cases, given that no linguistic tag was found to have a strong influence on the news, then the use of external knowledge would be necessary to detect the fake news.

- The most representative and common examples that result in the news items being classified as reliable are news fragments that present specific dates, places and subjects, as well as facts reported in an unbiased manner. The following are a few examples:
  - Example: **el pasado 25 de enero las autoridades chinas** anunciaban la construcción de dos hospitales para atender a los pacientes infectados con el coronavirus. (*On January 25, the Chinese authorities announced the construction of two hospitals to care for patients infected with the coronavirus.*)
  - Example: **A las 7 en punto de la mañana del 25 de enero**, más de **500 trabajadores de la construcción** *y* **más de 10 vehículos de maquinaria de construcción** aparecieron en el **Centro Médico Regional Dabieshan**, para transformarlo en el **Hospital Xiaotangshan**, en tan sólo **48 horas**, con capacidad para **más de 1.000 camas** (*At 7 o'clock on the morning of January 25, more than 500 construction workers and more than 10 construction machinery vehicles appeared at Dabieshan Regional Medical Center to transform it into Xiaotangshan Hospital in just 48 h, with a capacity of more than 1,000 beds.*)

  As can be seen in these examples, there is a large amount of concrete data that could be submitted to fact-checking, which gives a high level of reliability at the content level. However, after fact-checking, the data provided is shown not to be true. There are denials by verification agencies for both examples.[16][17]
- In some cases we also detect the presence of quotations in which the author's position is not shown, but the quotations are used to expand the information and corresponding evidence, which is a sign of neutrality and therefore of reliability.
  - Example: **"Los virus se vuelven más violentos en temas de virulencia, no necesariamente más letales"**, dijo Quintero en BLU Radio. (***"'Viruses become more aggressive in terms of virulence, not necessarily more lethal"**, Quintero told BLU Radio.*).

From the study, it can be concluded that for (U-T), despite finding information that was inaccurate and subjective, the news turned out to be true, since current journalistic techniques tend to try to hook the reader's attention in one sentence in the hope that the entire article will be read. Therefore, unreliable news is not always false.

In fact, fake news is increasingly being written more professionally, mixing true and false data with the intention of being more difficult for a reader to detect the false elements. This means that there is a strong argument for checking this type of news with external knowledge of the world from reliable sources.

Therefore, this work supports the claim that the automatic analysis of content in terms of reliability helps in the task of detecting fake news.

[16] http://bit.ly/3FlN9TJ

[17] https://bit.ly/3UkQ9E8

Furthermore, after the in-depth analysis performed in this research as to the relationship between reliability and veracity, reliability detection was shown to contribute greatly to detecting disinformation. However, results could be improved by developing a hybrid approach to determine veracity, which combines our content approach with a context approach (world-knowledge checking), being an effective solution for the disinformation detection task.

8. Conclusions and further work

The main novelty of this work is the design and implementation of a methodology to simplify the annotation task of semantically complex datasets, addressing the challenge of minimizing human effort and maximizing human feedback for annotated resources construction. The methodology exploits summarization and HITL techniques, so the application of the methodology results in an twofold improvement of the annotation task: automatically selecting the most relevant information of the news items and providing pre-annotated suggestions with a high degree of certainty. The methodology is applied in the disinformation context, but it could be easily adapted to whatever complex annotation task following the two levels and optimizing any annotation procedure.

In our case, within the disinformation task, a specific annotation scheme that focuses on unreliable news content (RUN-AS) is used and a semi-automatic annotation of a dataset is generated, following the designed methodology. The existing Spanish news' dataset with 9 different topics was used as a basis, whose news items were originally annotated with a veracity value (true/false) for the whole news item [58]. A subset of about 50% of this dataset was annotated using the proposed methodology.

The experiments conducted show that automatic pre-annotation and summarization reduce the annotation time by almost 50% and the annotated information by around 42%. Moreover, the use of RUN-AS features in the reliability detection task significantly increases the results with respect to a baseline without these features. The most significant case is found in the application of the SVM approach which goes from F_1m = 0.38 to 0.95. Moreover, the reliability detection model was shown to be valid for fake news detection, obtaining results in line with those obtained by other state-of-the-art fake news detection systems (0.78 Accuracy and 0.78 F_1m). Therefore, from these results it can be concluded that the proposed semi-automatic annotation methodology is highly suitable for the generation of efficient and effective datasets for training a ML or DL system, reducing time, effort and resources needed to generate the necessary annotated examples. Also, the proposed semi-automatic annotation is especially useful in cases such as RUN-AS, which is a fine-grained annotation with a high complexity load. Finally, an additional conclusion to this research is that reliability detection was shown to make a significant contribution to disinformation detection.

In future lines of this research the methodology will be applied to other languages and other complex annotation problems in different domains, by experimenting with different summarization approaches or replacing the pre-annotation module to the specific annotation problem. Furthermore, after the analysis of the inter-annotator agreement, the necessary actions will be taken to resolve misinterpretations of the guidelines or the instructions given for the annotation of those labels that have been detected as more complex. Finally, considering the benefits of detecting reliability in the fake news detection task, further work will involve developing a hybrid approach, whereby content (reliability) is combined with context via accessing external sources of world knowledge, such as fact-checking websites or scientific evidences. Furthermore, exploiting the benefits of the fine-grained annotation proposed by RUN-AS so as to contribute to the explainability of the predictions obtained will be explored [72].

CRediT authorship contribution statement

Alba Bonet-Jover: Formal analysis, Investigation, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing, Data curation. **Robiert Sepúlveda-Torres:** Data curation, Investigation, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Estela Saquete:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Patricio Martínez-Barco:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing 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.

Data availability

I have shared the links on the manuscript.

Acknowledgments

This research work is funded by MCIN/AEI/, Spain 10.13039/5011000 11033 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, Spain through the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/ 2021/21), and the grant ACIF/2020/177.

References

[1] V.L. Rubin, Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions, J. Doc. 75 (5) (2019) 1013–1034.
[2] D. Fallis, The varieties of disinformation, Philos. Inf. Qual. 358 (2014) 135–161.
[3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (6380) (2018) 1146–1151.
[4] S. Feng, R. Banerjee, Y. Choi, Syntactic stylometry for deception detection, in: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8–14, 2012, Jeju Island, Korea - Volume 2: Short Papers, The Association for Computer Linguistics, 2012, pp. 171–175.
[5] R. Mihalcea, C. Strapparava, The Lie detector: Explorations in the automatic recognition of deceptive language, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, 2009, pp. 309–312.
[6] A.X. Zhang, A. Ranganathan, S.E. Metz, S. Appling, C.M. Sehat, N. Gilmore, N.B. Adams, E. Vincent, J. Lee, M. Robbins, et al., A structured response to misinformation: Defining and annotating credibility indicators in news articles, in: Companion Proceedings of the the Web Conference 2018, 2018, pp. 603–612.
[7] R.M. Monarch, Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI, Simon and Schuster, 2021.
[8] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, 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, Association for Computational Linguistics, 2012, pp. 102–107.
[9] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, 2018, pp. 3391–3401.
[10] F.K.A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber, M. Farah, Fa-kes: A fake news dataset around the syrian war, in: Proceedings of the International AAAI Conference on Web and Social Media, 2019, Vol. 13, 2019, pp. 573–582.
[11] R.M. Silva, R.L. Santos, T.A. Almeida, T.A. Pardo, Towards automatically filtering fake news in portuguese, Expert Syst. Appl. 146 (2020) 113199.
[12] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (3) (2020) 171–188.
[13] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Springer, 2021, pp. 21–29.
[14] W.S. Paka, R. Bansal, A. Kaushik, S. Sengupta, T. Chakraborty, Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection, Appl. Soft Comput. 107 (2021) 107393.
[15] G.K. Shahi, J.M. Struß, T. Mandl, Overview of the CLEF-2021 CheckThat! lab: Task 3 on fake news detection., in: CLEF (Working Notes), 2021, pp. 406–423.
[16] W.Y. Wang, Liar, liar pants on fire: A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, 2017, pp. 422–426.
[17] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, 2014, pp. 18–22.
[18] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819.
[19] A. Khalil, M. Jarrah, M. Aldwairi, Y. Jararweh, Detecting arabic fake news using machine learning, in: 2021 Second International Conference on Intelligent Data Science Technologies and Applications, IDSTA, IEEE, 2021, pp. 171–177.
[20] S. Dhoju, M. Main Uddin Rony, M. Ashad Kabir, N. Hassan, Differences in health news from reliable and unreliable media, in: Companion Proceedings of the 2019 World Wide Web Conference, 2019, pp. 981–987.
[21] R. Assaf, M. Saheb, Dataset for arabic fake news, in: 2021 IEEE 15th International Conference on Application of Information and Communication Technologies, AICT, IEEE, 2021, pp. 1–4.
[22] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J.J.M. Escobar, Detection of fake news in a new corpus for the Spanish language, J. Intell. Fuzzy Systems 36 (5) (2019) 4869–4876.
[23] H. Gómez-Adorno, J.P. Posadas-Durán, G.B. Enguix, C.P. Capetillo, Overview of fakedes at iberlef 2021: Fake news detection in Spanish shared task, Procesamiento Lenguaje Nat. 67 (2021) 223–231.
[24] S.A. Brown, The Effects of Explicit Main Idea and Summarization Instruction on Reading Comprehension of Expository Text for Alternative High School Students (Ph.D. thesis), Utah State University, 2018.
[25] J. Engelen, G. Camp, J. van de Pol, A. de Bruin, Teachers' monitoring of students' text comprehension: can students' keywords and summaries improve teachers' judgment accuracy? Metacognition Learn. 13 (3) (2018) 287–307.
[26] X.G. Lin, S.-E. Jhang, D. Dong, Investigating the effects of text summarization on linguistic quality of argumentative writing, New Korean J. Engl. Lang. Lit. 60 (4) (2018) 245–268.
[27] J.P. Barreiro, Improving Reading Comprehension of Narrative Texts through Summaries (Ph.D. thesis), Universidad Casa Grande, 2019.
[28] R. Dijkman, A. Wilbik, Linguistic summarization of event logs - A practical approach, Inf. Syst. 67 (2017) 114–125.
[29] J. Petkovic, V. Welch, M. Jacob, M. Yoganathan, A.P. Ayala, H. Cunningham, P. Tugwell, The effectiveness of evidence summaries on health policymakers and health system managers use of evidence from systematic reviews: A systematic review, Implement. Sci. 11 (2016).
[30] L. Hartling, A. Gates, J. Pillay, M. Nuspl, A. Newton, Development and usability testing of epc evidence review dissemination summaries for health systems decisionmakers, Methods Research Report. Technical Report EHC027-EF, Agency for Healthcare Research and Quality (US), 2018.
[31] Y. Liu, X. Song, S.-F. Chen, Long story short: finding health advice with informative summaries on health social media, Aslib J. Inf. Manag. 71 (6) (2019) 821–840.
[32] W.S. El-Kassas, C.R. Salama, A.A. Rafea, H.K. Mohamed, Automatic text summarization: A comprehensive survey, Expert Syst. Appl. 165 (2021) 113679.
[33] E. Lloret, M. Palomar, Text summarisation in progress: a literature review, Artif. Intell. Rev. 37 (1) (2012) 1–41.

[34] N. Moratanch, S. Chitrakala, A survey on extractive text summarization, in: 2017 International Conference on Computer, Communication and Signal Processing, ICCCSP, 2017, pp. 1–6, http://dx.doi.org/10.1109/ICCCSP.2017.7944061.
[35] D. Jani, N. Patel, H. Yadav, S. Suthar, S. Patel, A concise review on automatic text summarization, in: J. Nayak, H. Behera, B. Naik, S. Vimal, D. Pelusi (Eds.), Computational Intelligence in Data Mining, Springer Nature Singapore, 2022, pp. 523–536.
[36] M. Kirmani, N. Manzoor Hakak, M. Mohd, M. Mohd, Hybrid text summarization: A survey, in: K. Ray, T.K. Sharma, S. Rawat, R.K. Saini, A. Bandyopadhyay (Eds.), Soft Computing: Theories and Applications, Springer Singapore, 2019, pp. 63–73.
[37] E. Okoro, B. Abara, A. Umagba, A. Ajonye, Z. Isa, A hybrid approach to fake news detection on social media, Niger. J. Technol. 37 (2) (2018) 454–462.
[38] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artif. Intell. Rev. (2022) 1–50.
[39] G. Demartini, S. Mizzaro, D. Spina, Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities, IEEE Data Eng. Bull. 43 (3) (2020) 65–74.
[40] M. Fanton, H. Bonaldi, S.S. Tekiroglu, M. Guerini, Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech, 2021, arXiv abs/2107.08720.
[41] H. Cañizares-Díaz, A. Piad-Morffis, S. Estevez-Velarde, Y. Gutiérrez, Y.A. Cruz, A. Montoyo, R. Muñoz, Active learning for assisted corpus construction: A case study in knowledge discovery from biomedical text, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2021, 2021, pp. 216–225.
[42] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Gener. Comput. Syst. 135 (2022) 364–381.
[43] K. Tomanek, J. Wermter, U. Hahn, An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL, 2007, pp. 486–495.
[44] M. Kholghi, L. Sitbon, G. Zuccon, A. Nguyen, Active learning: a step towards automating medical concept extraction, J. Am. Med. Inform. Assoc. 23 (2) (2016) 289–296.
[45] L. Benedikt, C. Joshi, L. Nolan, R. Henstra-Hill, L. Shaw, S. Hook, Human-in-the-loop AI in government: A case study, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20, Association for Computing Machinery, 2020, pp. 488–497.
[46] S. Budd, E.C. Robinson, B. Kainz, A survey on active learning and human-in-the-loop deep learning for medical image analysis, Med. Image Anal. 71 (2021) 102062.
[47] W. Jung, F. Jazizadeh, Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions, Appl. Energy 239 (2019) 1471–1508.
[48] A.M. Daniel, Human-in-the-loop disinformation detection: Stance, sentiment, or something else? 2021, CoRR abs/2111.05139.
[49] H. Zhang, H. Liu, Visualizing structural "inverted pyramids" in English news discourse across levels, Text Talk 36 (1) (2016) 89–110.
[50] W. Wang, D. Zhao, L. Zou, D. Wang, W. Zheng, Extracting 5w1h event semantic elements from Chinese online news, in: International Conference on Web-Age Information Management, Springer, 2010, pp. 644–655.
[51] H. Rashkin, E. Choi, J.Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2931–2937.
[52] B. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, in: Proceedings of the International AAAI Conference on Web and Social Media, 2017, pp. 759–766, http://dx.doi.org/10.1609/icwsm.v11i1.14976.
[53] S. Mottola, Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español, Discurso Sociedad 14 (3) (2020) 683–706.
[54] R. Mihalcea, P. Tarau, TextRank: Bringing order into text, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2004, pp. 404–411.
[55] J. Canete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR, Vol. 2020, 2020.
[56] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 2383–2392.
[57] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, 2020, arXiv abs/2107.08720.
[58] J. Posadas-Durán, H. Gomez-Adorno, G. Sidorov, J. Escobar, Detection of fake news in a new corpus for the Spanish language, J. Intell. Fuzzy Systems 36 (5) (2019) 4868–4876.
[59] A. Bonet-Jover, A. Piad-Morffis, E. Saquete, P. Martínez-Barco, M.Á.G. Cumbreras, Exploiting discourse structure of traditional digital media to enhance automatic fake news detection, Expert Syst. Appl. 169 (2021) 114340.
[60] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37.
[61] M. Boguslav, K. Cohen, Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing, Stud. Health Technol. Inform. 245 (2017) 298–302.
[62] G. Hripcsak, A.S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, J. Am. Med. Inform. Assoc. 12 (3) (2005) 296–298.
[63] J. Legrand, R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, M. Smail, Y. Toussaint, A. Coulet, PGxCorpus, a manually annotated corpus for pharmacogenomics, Sci. Data 7 (2020) 3.
[64] V.L. Rubin, Y. Chen, N.J. Conroy, Deception detection for news: Three types of fakes, 2016.
[65] A. Altheneyan, A. Alhadlaq, Big data ML-based fake news detection using distributed learning, IEEE Access 11 (2023) 29447–29463.
[66] K. Ma, C. Tang, W. Zhang, B. Cui, K. Ji, Z. Chen, A. Abraham, DC-CNN: Dual-channel convolutional neural networks with attention-pooling for fake news detection, Appl. Intell. 53 (7) (2023) 8354–8369.
[67] K. Li, HAHA at FakeDeS 2021: A fake news detection method based on TF-IDF and ensemble machine learning, in: IberLEF@ SEPLN, 2021, pp. 630–638.
[68] T. Jiang, J.P. Li, A.U. Haq, A. Saboor, A. Ali, A novel stacking approach for accurate detection of fake news, IEEE Access 9 (2021) 22626–22639.
[69] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C.P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Maria: Spanish language models, 2021, arXiv preprint arXiv:2107.07253.
[70] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, CoRR abs/1907.11692.
[71] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, P.M. Barco, Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation, Procesamiento Lenguaje Nat. 70 (2023) 15–26.
[72] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, in: AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.