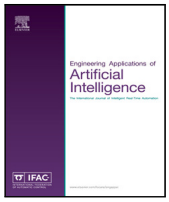
Engineering Applications of Artificial Intelligence 126 (2023) 107152

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Applying Human-in-the-Loop to construct a dataset for determining content reliability to combat fake news

Alba Bonet-Jover [a,*], Robiert Sepúlveda-Torres [a], Estela Saquete [a], Patricio Martínez-Barco [a], Alejandro Piad-Morffis [b], Suilan Estevez-Velarde [b]

[a] Department of Software and Computing Systems, University of Alicante, Spain
[b] School of Math and Computer Science, University of Havana, Cuba

ARTICLE INFO

Keywords:
Natural language processing
Fake news detection
Assisted annotation
Dataset construction
Human-in-the-Loop Artificial Intelligence
Active learning

ABSTRACT

Annotated corpora are indispensable tools to train computational models in Natural Language Processing. However, in the case of more complex semantic annotation processes, it is a costly, arduous, and time-consuming task, resulting in a shortage of resources to train Machine Learning and Deep Learning algorithms. In consideration, this work proposes a methodology, based on the human-in-the-loop paradigm, for semi-automatic annotation of complex tasks. This methodology is applied in the construction of a reliability dataset of Spanish news so as to combat disinformation and fake news. We obtain a high quality resource by implementing the proposed methodology for semi-automatic annotation, increasing annotator efficacy and speed, with fewer examples. The methodology consists of three incremental phases and results in the construction of the RUN dataset. The annotation quality of the resource was evaluated through time-reduction (annotation time reduction of almost 64% with respect to the fully manual annotation), annotation quality (measuring consistency of annotation and inter-annotator agreement), and performance by training a model with RUN semi-automatic dataset (Accuracy 95% F1 95%), validating the suitability of the proposal.

1. Introduction

Disinformation is considered a type of "information disorder" (Wardle et al., 2018) whereby false information is deliberately created or disseminated with the express intention of causing harm. This includes fake news and hoaxes. Despite the term fake news being widely known, this research adopts the term disinformation which embraces a more comprehensive concept of the global problem (Ireton and Posetti, 2018). Although disinformation has always existed, the real threat today is its rapid dissemination via digital media that can end up generating public health, social or ideological problems. Technological progress has resulted in a more connected world but along with greater connectivity comes the potential for its misuse (Shu et al., 2020). This rapid and viral spread of huge amounts of information makes automatic detection necessary because it is impossible to manually process this volume of data. Moreover, existing algorithms for automatic detection require the intervention of human experts as the system needs annotated examples to learn from expert feedback as well as to justify the decision taken. Obtaining this enormous amount of annotated data is a highly costly task, both in terms of resources and time.

Automatic disinformation detection is a complex task to be resolved from an engineering point of view, and the research community is approaching this task from different perspectives (Saquete et al., 2020), such as stance detection, polarization, credibility, or automated fact-checking. Our research is grounded in the context of credibility, focusing on automatically detecting reliability of documents, by means of Artificial Intelligence (AI) and Natural Language Processing (NLP). Our approach is a content-based approach, which only uses the content of the document and does not require external knowledge information to decide this reliability. This is a first screening step towards a more complex task of determining the veracity of a document, for which it is mandatory to check the information also with external knowledge. In this case, veracity detection is a step further than the scope of this work.

One of the main challenges in our research is the scarcity of training data. Training corpora created by human experts are essential in NLP.

* 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), apiad@matcom.uh.cu (A. Piad-Morffis), sestevez@matcom.uh.cu (S. Estevez-Velarde).

https://doi.org/10.1016/j.engappai.2023.107152
Received 28 July 2022; Received in revised form 4 July 2023; Accepted 11 September 2023

When a problem is approached from the AI perspective, either with Machine Learning (ML) or Deep Learning (DL) techniques, large and costly amount of instances of human feedback are required to construct the datasets that will be used to train and evaluate the systems that will be in charge of solving the problem (Stenetorp et al., 2012). With current pre-trained models (i.e. based on Transformer architecture (Vaswani et al., 2017)), the number of examples required is smaller than in classical approaches but still very expensive.

Building efficient datasets is a complex task as dataset annotations can have different degrees of difficulty. This may imply not only a time cost but also the need for a high level of expertise in a given annotation. An efficient dataset would be one that can be created as quickly and inexpensively as possible and also includes the most appropriate annotated examples to assist learning for problem-solving.

Our challenge is to create tailor-made quality datasets, selected according to specific criteria, that increase, or at least maintain, accuracy while saving time and effort. This would result in larger and more efficient datasets that combine both automatic and manual annotation. Moreover, the datasets need to be constantly updated at a minimum cost so that the tools derived from them do not become obsolete.

Hence, the overall objective of this work is to implement a novel methodology for semi-automatic dataset construction that will allow the efficient and effective generation of quality resources. In our research, we will focus on news content reliability to support disinformation detection, but the methodology could be easily adapted and applied to any semantically complex annotation task.

To address the overall objective, we focus on a paradigm called Human-in-the-Loop. The Human-in-the-Loop (HITL) paradigm is an extensive area of research that covers the intersection of computer science, cognitive science, and psychology, and of course, it is being applied in the AI area. Specifically, HITL machine learning is a set of strategies for combining human and machine intelligence in applications that use AI (Monarch, 2021).

As a result, the following contributions to this research area have been provided:

- The design and implementation of an innovative HITL-based methodology for semi-automatic annotation in complex annotation tasks thereby assisting annotators and optimizing resources and performance, while facilitating the periodic updating of the language models used by the tools.
- The creation of an efficiently annotated dataset by applying the proposed methodology, which is a fundamental requirement for AI and NLP tasks. In this work, the dataset is applied to disinformation detection and, more specifically, reliable and unreliable information in news articles in Spanish.
- The evaluation of the quality and the benefits of applying this type of HITL-based methodology for the semi-automatic construction of datasets. The performance is determined in terms of a balance between time-effort consumption, annotation quality of the dataset, and accuracy achievement.
- Making available to the research community the semi-automatic dataset generated,[1] once the validity of the generated dataset has been corroborated.

This paper is structured as follows: Section 2 presents an overview of the most relevant scientific literature concerning disinformation datasets, human-in-the-loop AI, and corpus construction methodologies; Section 3 details RUN-AS, the annotation guideline created for our research; Section 4 introduces the methodology for semi-automatic annotation of datasets; Section 5 presents the specific implementation of the methodology; Section 6 describes the evaluation framework and discussion; and finally Section 7 presents the conclusions of this research and future work.

[1] Available at https://gplsi.dlsi.ua.es/resources/NewsReliabilityAnnotation

2. Background

Our research is based on the application of AI for disinformation detection where the dataset becomes the cornerstone. In this context, we face two challenges. Firstly is the lack of high-quality datasets with labeled examples in Spanish on which statistical models for automatic disinformation detection can be trained. Secondly is the time and effort required to obtain the examples. As stated by Alex et al. (2010), very little work has been done to obtain better annotation methods that maximize both the quality and quantity of the annotated data.

This section presents the state of the art related to disinformation datasets (Section 2.1), followed by the literature regarding the human-in-the-loop concept (Section 2.2), and finally the different methodologies usually adopted in corpus construction in NLP (Section 2.3).

2.1. Disinformation corpora

According to the literature consulted, most of the corpora created to address disinformation or fake news detection use a binary classification, by categorizing news as Fake or True (Salem et al., 2019; Silva et al., 2020). Others, such as those focused on fact-checking tasks, use a fine-grained scale of labels covering several veracity degrees (Wang, 2017; Vlachos and Riedel, 2014). However, in all cases, the annotation is global for the whole document and the veracity or reliability of the different parts of the document are not considered. This single global classification of news, whether with binary or multiple values, depends on external knowledge, such as fact-checking platforms. Few datasets use a reliability classification and usually this classification is applied on the basis of the source's credibility (Dhoju et al., 2019) and not of purely textual or linguistic characteristics (Assaf and Saheb, 2021).

To the authors' knowledge, corpora that address the disinformation task in Spanish (Posadas-Durán et al., 2019) are scarce, since they are usually released in English. Therefore, we aim to create a Spanish resource to train in this task. Furthermore, all the disinformation datasets found in the literature were created entirely manually, thus as explained in next section we introduce more efficient ways to create datasets.

2.2. Human-in-the-loop machine learning

The use of supervised learning is approximately 90% of today's machine learning-based applications, i.e. they learn based on examples created by humans. As stated by Okoro et al. (2018), there is a need for a hybrid model solution that combines the efforts of both humans and machines. Due to the complexity of our semantic annotation proposal and given that researchers spend more time generating data than building machine learning models (Monarch, 2021), we focus on the HITL methodologies to increase the efficacy of our work. HITL is an umbrella term for defining the new types of interactions between humans and machine learning algorithms (Mosqueira-Rey et al., 2022).

HITL-AI are systems that continuously improve because of human input, addressing the limitations of previous AI solutions and bridging the gap between machines and humans. These systems aim at leveraging the ability of AI to scale the processing to very large amounts of data while relying on human intelligence to perform very complex tasks, such in the case of natural language understanding (Demartini et al., 2020). The HITL methodology is being used in several studies to increase efficiency in data collection, such as in the cases of Fanton et al. (2021) and Cañizares-Díaz et al. (2021), since the continuous executive loop contributes to higher accuracy and stronger robustness of the systems. Fanton et al. (2021) proposed a novel human-in-the-loop data collection methodology in which a generative language model is refined iteratively by using its own data from the previous loops to generate new training samples that experts review and/or post-edit. Cañizares-Díaz et al. (2021) applies the HITL approach active learning to reduce the human effort required during the annotation of natural language

corpora composed of entities and semantic relations. The approach assists human annotators by intelligently selecting the most informative sentences to annotate and then pre-annotating them with a few highly accurate entities and semantic relations. Finally, a survey of existing works on human-in-the-loop from a data perspective are presented at Wu et al. (2022), summarizing major approaches in the field along with their technical strengths/weaknesses, and classifying them into three main categories: (i) the work of improving model performance from data processing; (ii) the work of improving model performance through interventional model training; and (iii) the design of the system independent human in-the-loop.

One of the principles of HITL is to assist human tasks with machine learning to increase efficiency. In line with this, our work builds a semi-automatic annotated dataset, but HITL is used in many parts of ML cycle, from sampling unlabeled data to updating the model.

Depending on who is in control of the learning process, we can identify different approaches to HITL-ML (Mosqueira-Rey et al., 2022):

- **Active Learning (AL)**: Active learning is a very extended HITL strategy used when obtaining labeled data demands a large amount of time or money, since AL aims at selecting examples with high utility for the model (Tomanek et al., 2007) thereby increasing the performance of the learning model while reducing the amount of annotated data required (Kholghi et al., 2016). In AL methods, labels are collected from humans, fed back to a supervised learning model, and used to decide which data items humans should label next (Spina et al., 2015). AL is applied in several ML tasks such as object detection, semantic segmentation, sequence labeling or language generation (Monarch, 2021). In our work, AL is focused on disinformation detection and it enables us to optimize performance with fewer but better chosen news documents to incorporate in our training set. An important aspect of AL is the iterative process, since it allows the retraining of human feedback, which in turn enables the system to improve in terms of accuracy.
  Two broad active learning sampling strategies are (Monarch, 2021):
  - Uncertainty sampling. This is the set of strategies for identifying unlabeled items that are near a decision boundary in the current machine learning model.
  - Diversity sampling. This is the set of strategies for identifying unlabeled items that are underrepresented or unknown for the machine learning model in its current state. The goal of this sampling is to target new, unusual and underrepresented items for annotation to give the model a more complete picture of the problem space.
- **Interactive Machine Learning (IML)**: IML is an active machine learning technique in which models are designed and implemented with humans in the loop (Fails and Olsen, 2003; Wondimu et al., 2022). When using machine learning in an interactive design setting, feature selection must be automatic rather than manual and classifier training-time must be relatively fast. There is a closer interaction between users and learning systems, with people interactively supplying information in a more focused, frequent, and incremental way compared to traditional machine learning (Amershi et al., 2014). Ramos et al. (2020) defines IML as "the process in which a person (or people) engages with a learning algorithm, in an interactive loop, to generate useful artifacts". The authors describe these artifacts as data, insights about data, or machine-learned models. The difference between AL and IML relies more on who has the control of the learning process and not on the interactivity of the approach. While in AL the model retains the control and uses the human as an oracle, in IML there is a closer interaction between users and learning systems, so the control is shared. AL focuses on building better models in an algorithm-centered evaluation, but in IML systems human factors have to be taken into account, so there is also a human-centered evaluation, focusing on the utility and effectiveness of the application for end-users. AL is considered the basis for IML (Mosqueira-Rey et al., 2022).
- **Machine Teaching (MT)**: MT describes the idea of a teacher who teaches an ML model to an ML algorithm. In MT, human domain experts have control over the learning process by delimiting the knowledge that they intend to transfer to the machine learning model (Ramos et al., 2020; Simard et al., 2017). Even though the MT paradigm is quite different in nature from the other paradigms described in this paper (and represents an alternative to them) there are many common factors. Over time, the process has become iterative and incremental. Occasionally, it has been inspired by other approaches, such as active learning. MT has, at times, ended up obtaining results that are comparable to other techniques, such as curriculum learning.[2]

Beyond these well-known approaches, the HITL paradigm encompasses all those strategies that include two 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 our proposal, both goals are involved in the design of a methodology to create a semi-automatic annotation platform that enables an increase in the amount of annotated data, reaching the target accuracy more quickly and easily.

HITL-ML has been successfully applied in a variety of areas such as government (Benedikt et al., 2020), medicine (Budd et al., 2021), and energy (Jung and Jazizadeh, 2019). More specifically, as for applying HITL to dis- and mis-information detection, some works are key. Demartini et al. (2020) 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. Additionally, Daniel (2021) proposed a human-AI hybrid disinformation detection system performing as follows: The human user identifies a topic or claim about which they believe disinformation will be found, and seeks to learn more about what is being said and (possibly) who is saying it. The machine learning algorithm, taking in the topic/claim and large amounts of text scraped from the internet, is able to, by using the appropriate approval/disapproval relationships (as decided by the user), separate the text into two groups: relevant disinformation, and everything else. Closing the loop, the human can determine if it satisfies their interests or not, and if necessary, revise the search and run the process again. The paper aimed to determine which of stance, sentiment (or something else) techniques is best suited for use in this human-in-the-loop disinformation detection. The author indicates that the results obtained are not conclusive but data suggest that sentiment analysis algorithms outperform the stance detection algorithms. However, sentiment analysis and stance methods should be studied further for this purpose. These two papers mentioned above, although they do address the issue of disinformation using HITL techniques, are different perspectives to the one presented in this paper.

2.3. Corpora construction methodologies

The design, creation and annotation of a corpus is an essential task in the development of tools and datasets in NLP but, as stated by Stenetorp et al. (2012), "annotation is also one of the most time-consuming and financially costly components of many NLP research efforts". Nowadays, the number of labeled datasets available for training purposes is low and data collection is one of the challenges in

[2] Curriculum learning (CL) is a training strategy that trains a machine learning model from easier data to harder data, which imitates the meaningful learning order in human curricula

deception research due to the scarce availability of such datasets (Saquete et al., 2020). This scarcity is due to the time and cost that the annotation task requires because annotating and compiling a corpus demands effort, time, consistency, and human expertise. This subject is at the forefront of NLP research and particularly of disinformation detection research, since "the development of new resources such as annotated corpora can help to increase the performance of automatic methods aiming at detecting this kind of news" (Posadas-Durán et al., 2019).

According to the literature consulted, corpus construction in NLP can be approached via several methodologies. Even if there are cases in which both the compilation and the annotation tasks are completely automated (Abacha et al., 2015) or carried out manually (Evrard et al., 2020), most of the corpora released for the disinformation task follow semi-automatic, but not intelligent, methodologies. In the semi-automatic approach, data collection is mostly carried out in an automatic way via social media, fact-checking websites APIs, and web crawling or web scraping, whereas the annotation task is mostly carried out manually by experts, such as the corpora introduced by Shahi and Nandini (2020) and Wang (2017). Even if the manual annotation allows quality examples to be obtained, created and verified by experts, it is an arduous process that leads to small-size resources that require more time to achieve the desired goal.

Another type of methodology is crowdsourcing, in which both compilation and annotation can be automatic or manual, such as those introduced by Mitra and Gilbert (2015), Färber et al. (2020), Pérez-Rosas and Mihalcea (2015). This practice enables the bulk outsourcing of multiple labeling tasks, typically with low overall cost and fast completion (Hsueh et al., 2009). It facilitates the creation of larger training datasets, but the quality is often lower than those corpora developed especially by teams of experts working in the same field and cooperating in the same research group.

Besides semi-automatic and crowdsourced corpora, there is increasing interest in applying supervised or semi-supervised learning to build corpora (Feller et al., 2018). Fairly extensive research grew out of the Text REtrieval Conference (TREC)[3] that resulted in enhancing the fairness and efficiency of the annotation tasks. The TREC's task was based on judging the document relevance, not token-level annotations such as that done in the present work, but the approaches presented constitute a very important basis for dataset generation methodologies (Voorhees, 2018; Vu and Gallinari, 2006). Furthermore, applying Human-in-the-loop strategies, and especially Active Learning, to obtain datasets more efficiently is not new (Olsson, 2009). These approaches enable the creation of quality resources supervised by human experts, thus obtaining a considerable corpus through automation while keeping the quality of the human process. In this case, the system makes decisions in an automatic way but under the supervision of the expert, who corrects, validates or refutes those decisions (Cañizares-Díaz et al., 2021; Rahman et al., 2020). Most of the research in the literature on facilitating the annotation of entities in datasets, through supervised learning or human in the loop, is applied in the medical domain Kholghi et al. (2017), Tchoua et al. (2019), Settles et al. (2007).

Therefore, considering the task of disinformation and fake news detection and taking into account that this task requires evidences to justify why a certain decision has been made about the veracity of a news item, this implies a finer-grained annotation that allows for the explainability of the model and the obtained veracity classification, instead of a unique veracity value of the whole document as the state-of-the-art works do. But, at the same time, this finer-grained annotation also makes the work of annotating the datasets difficult and costly, so it is necessary to find a methodology that allows the construction of these datasets in an efficient and effective way. The proposals of HITL will collaborate in this task of improving efficiency and they have been successfully tested in other fields. The novelty of this work is to propose a methodology, which can be generalized to any complex annotation task, and which facilitates the creation of these complex, high quality resources.

To the best of the authors' knowledge, none of the works presented in the literature addresses the annotation process for reliability detection within the disinformation context in Spanish by means of the HITL paradigm.

Although not the main aim of this work, Section 3 explains the peculiarities of the annotation scheme proposed for the disinformation dataset used. The rationale being so that the construction process can be better understood, given its complexity on account of its high semantic and linguistic load.

3. RUN-AS annotation proposal

Our work is focused on the annotation of news collected from digital newspapers in Spanish and belonging to different domains, and it is based on two well-known journalistic techniques: the Inverted Pyramid and the 5W1H.

Regarding the journalistic structure, well-built news using the inverted pyramid tends to present five common parts, placed in order of relevance (Zhang and Liu, 2016), which are TITLE, SUBTITLE, LEAD, BODY and CONCLUSION. According to Thomson et al. (2008), neutrality and the inverted pyramid structure are distinctive features of hard news.

In terms of content, well-built articles present semantic information represented through a journalistic technique known as the 5W1H, whose elements "clearly describe key information of news in an explicit manner" (Zhang et al., 2019). The technique consists of answering six key questions: 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" (Wang et al., 2010). Hamborg et al. (2018) explains that journalists typically answer these questions for describing the main event of a news story and they are answered within the first few sentences of a news article. These studies point out that the reliability of the information lies in the clearly identifiable existence of these items, as well as in the way they are expressed. This is why our annotation for measuring reliability is based on these two journalistic practices.

A fine-grained annotation guideline called RUN-AS (Reliable and Unreliable News Annotation Scheme)[4] has been designed to train our dataset specifically created for the reliability detection of news. The novelty of this annotation scheme lies in the reliability classification based on 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 enable the detection of suspicious information). We propose a complex semantic annotation that is based on a multi-level annotation, two journalistic techniques, and an in-depth linguistic analysis. For this reason, our annotation required expert linguistic annotators as well as technical experts for building the algorithms and models. An example of this annotation is shown in Fig. 1.

To the authors' knowledge, current datasets focus on determining a global and single veracity value of the news items. However, our proposal enables the annotation of essential content within a news item and assigns a reliability classification based on a purely textual, linguistic and semantic analysis that takes into account several elements such as vagueness, subjectivity, lack of evidence or emotionally charged content that influences reader opinions and feelings (Zhang et al., 2019). The complete reliability criteria based on accuracy and neutrality concepts are fully defined in Bonet-Jover et al. (2023). In

[3] https://trec.nist.gov/overview.html

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

<TITLE><WHO>Several experts</WHO> <WHAT>state that lemon can save your life</WHAT></TITLE>

<LEAD><WHEN>A few days ago</WHEN>, <WHO>several renowned experts</WHO> from <WHERE>California</WHERE> <WHAT>affirmed that lemon can save our lives</WHAT> <WHY>since it prevents and cures cancer.</WHY></LEAD>

<BODY><WHAT>Lemon has several properties</WHAT>, but according to <WHO>medical experts</WHO> <WHAT>this citrus fruit has been used</WHAT> <WHEN>for millions of years</WHEN> <HOW>with hot water</HOW>,<WHY>as drinking lemon infused water kills cancer cells in our body and creates a protective shield that prevents future tumours [...]</WHY></BODY

Fig. 1. Example of part of the structure (Inverted Pyramid) and content (5W1H) labels of a news item.

our annotation guideline, the emotional charge is being marked in the Elements of Interest, especially with the labels KEY_EXPRESSIONS, author_stance of the QUOTE, and style of the TITLE. All this information is provided in the reference included.

For predicting the veracity of a news item, world knowledge is essential, but what our proposal aims to achieve is not to detect veracity, but rather the features that can be decisive in classifying a news item as reliable or unreliable, thereby providing support to users and journalists via useful information at first level text-only annotation. In our annotation procedure, one expert linguistic annotator was involved to ensure the coordination and harmonization of the annotation process as well as compliance with the annotation guideline previously described. News was annotated using Brat,[5] an intuitive and practical annotation tool.

4. Human-in-the-loop based methodology for semi-automatic annotation

This section presents the design of a HITL-based methodology to semi-automate the dataset construction task. In order to simplify compilation and annotation tasks, the HITL paradigm was used to gradually automate the tasks involved in the construction of the dataset. This minimizes the effort of the human participant in the annotation, and creates larger and less costly datasets. This methodology could be easily adapted to whatever complex annotation task optimizing any annotation procedure as we will discuss in Section 5.4.

The methodology proposed here was performed in three phases consisting of gradually integrating automation into the annotation process and observing the changes compared to the fully manual annotation. The news items were annotated in small batches and following different strategies in each phase, but keeping the same number of news items for each batch and always following the annotation scheme.

Prior to the main procedure, there was a data collection stage to source a large set of news items from various news and information providers.

Fig. 2 shows a high-level diagram of the three phases of the methodology based on HITL: Phase 1: manual compilation and annotation of the corpus; Phase 2: automated compilation and manual annotation; and, Phase 3: automated compilation and semi-automatic annotation.

4.1. Phase 1: Reporting on the manual compilation and annotation

In the first phase, news was compiled and annotated in an entirely manual way, as illustrated in step 1 of Fig. 2. Concerning the compilation task, a total of 40 news items in Spanish was collected from 9 sources. Then, news was manually annotated following the RUN-AS guideline (see Section 3).

This first phase was an arduous and slow process, since searching for news items one by one and annotating them from scratch was time-consuming.

The result of this phase was the first version of the dataset (in Fig. 2 green cylinder), obtaining labeled news that were used as input for Phase 2.

4.2. Phase 2: automated compilation and manual annotation

The second phase introduced a well-known HITL strategy to increase the productivity of the annotation process. Specifically, an active learning approach was implemented in this phase, where the human annotator interacted with a machine learning model that automatically selected the most informative documents to annotate. Active learning was chosen over other existing HITL techniques because it provided an easy and unobtrusive way to enhance the annotator's performance, without requiring additional training of annotators. In fact, annotators do not even need to know there is a machine learning model selecting the documents to annotate. This means that we can leverage annotators with previous experience in the domain even though they may not be technically savvy to participate in more complex HITL scenarios.

The process involved in this work is described next. Starting from a small batch of annotated news items (from Phase 1), a supervised model was trained and applied to a larger batch from an unlabeled news pool (orange cylinder). For each item, an informativeness metric was computed (see Section 5.1) based on a balance between model uncertainty and content diversity. All unlabeled news items were sorted by this score, and a tentative list of suggestions was created, by interleaving news items from different sources. Thus, in this list, the most informative news items appeared first, taking into consideration content diversity. From this list of suggestions, an expert annotator filtered out those that did not follow the language, format, extension, or other semantic characteristics desired in the corpus. The final list consisted of the K most informative news items that fit all the desired criteria evaluated prior to annotation, and balanced in terms of the original sources. Finally, this batch of K news items was annotated and added to the training set (step 6 and step 7 in Fig. 2), and the whole active learning cycle was repeated. In Fig. 2, after steps 3, 4 and 5, the model selected the most appropriate news items to be annotated from the unlabeled news pool.

Specifically, a total of 4 batches was performed in this phase, and a total of 10 news items were selected in each batch. Thus, after Phase 2 finishes, 40 novel news items were added to the corpus. As explained, these news items were manually annotated by an expert annotator, but their selection, which was based on an active learning strategy, helped to guarantee a minimum level of diversity and consistency that would have been difficult to attain with a purely manual selection.

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

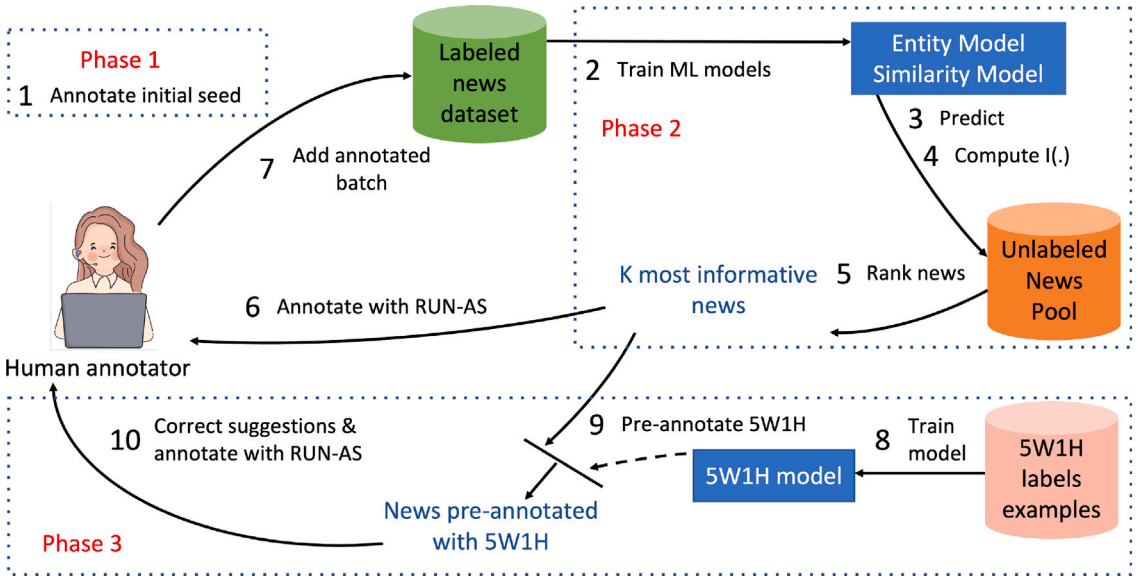


Fig. 2. High-level diagram of the three phases of the HITL-based methodology.

4.3. Phase 3: automated compilation and semi-automatic annotation

Finally, the third phase was an evolution of the second phase, with the aim of significantly improving annotation times. This phase performed a human–machine interaction consisting of the human reviewing and improving the automatic pre-annotation of the dataset, provided by an ML-assisted labeling system, in a machine-human-machine loop. This resulted in an improvement of the reliability detection classification by increasing the size of the training dataset. Furthermore, the new examples reviewed by the human served to re-train the ML-assisted labeling system.

The novelty here compared with Phase 2 is the pre-annotation carried out by the system to assist the expert. The purpose being for the annotator not to label from scratch, but only to revise and complete the pre-annotation done by the system. In this stage, the system only pre-annotated the news items with the second annotation level (5W1H labels) defined in RUN-AS because it is the more complex and time-consuming annotation level.

As seen in Fig. 2, human intervention remains important since it is necessary to check that the automatic selection of news and the pre-annotation proposed by the semi-automatic system meet the criteria of the dataset. In this sense, the loop presented in Phase 2 is extended (steps 8, 9, and 10). This phase used a 5W1H model, previously trained (step 8) with 5W1H labels examples (pink cylinder), to pre-annotate 5W1H labels (step 9) in the news selected. The last step in this phase is 10, where the pre-annotated items according to the 5W1H model were edited by the annotators, and the rest of RUN-AS annotation was added. Finally, a new annotated batch was added to the dataset (step 7) to conclude one loop. The new corrected examples were also used to re-train the 5W1H pre-annotation model.[6]

In this stage, 40 news items were initially annotated in order to keep the same number of annotated news as in the previous phases. However, after validating that the semi-automatic annotation accelerated the process (see Section 6.4), we decided to annotate another 50 news items (90 in total in this phase), in order to increase the dataset. A total of 9 batches were annotated.

[6] This loop is not indicated in the figure for clarity

5. Implementation of methodology

Following the conceptual definition of the methodology, a specific implementation was performed for the semi-automatic annotation of news content reliability. The specific implementation in this domain of Phase 2 and 3 are fully explained next. In the second phase, 4 batches were performed and in the third, 9 batches. All batches had 10 annotated news items. A proposal for the generalization of the methodology is presented at Section 5.4.

5.1. Phase 2 (active learning model)

The active learning model used in this phase was an implementation based on the proposal by Cañizares-Díaz et al. (2021) for entity and relation annotation. The original model consisted of two different classifiers, one for entity recognition and another for relation extraction. However, since our annotation scheme does not contain relations, only the entity classifier was used.

The model is based on a logistic regression classifier trained on token-level entity labels. Thus, a preprocessing of the annotated text was performed that transforms Brat-based annotation, which is defined at the text span level, into a sequence of annotated tokens. The logistic regression model was fed with token-level syntactic, semantic (extracted with spacy), and contextual features (i.e., the combined features of a small window of surrounding tokens).

To compute the informativeness of a news item, the trained model was executed on each sentence of the whole document, and the probability distributions of all possible labels in each token were stored. Based on this distribution, a token-level measure of entropy was computed as shown in Eq. (1), where $p^{(t)}_{label}$ is the probability associated with a specific label in token t.

$$H(t) = -\sum_{label} p^{(t)}_{label} \log p^{(t)}_{label} \tag{1}$$

The overall entropy of a document D was computed as the mean entropy of all its tokens $t \in D$ (see Eq. (2)), which corresponds to a standard interpretation of the annotation process as a stochastic process with independent decisions. This is a simplification since the labels of a specific token are often correlated with the labels of nearby tokens. However, this simplification makes the problem tractable and

requires no additional assumptions about the semantics of the annotation scheme, which makes it extendable to other annotations.

$$H(D) = \frac{1}{\|D\|} \sum_{t \in D} H(t) \tag{2}$$

Finally, a similarity factor $\sim (D_i, \mathbf{D})$ was defined between every new document D_i and the set of annotated documents $\mathbf{D}$. This similarity was computed as the mean dot-product similarity between document D_i and all documents already annotated in $\mathbf{D}$, based on their doc2vec representation obtained with the Python library `gensim` (see Eq. (3)). This similarity factor was used to decrease the informativeness of potential outliers, e.g., news items in other languages, or documents that are not news items but nevertheless were included in the unlabeled set during data collection.

$$sim(D_i, \mathbf{D}) = \frac{1}{\|\mathbf{D}\|} \sum_{D_j \in \mathbf{D}} \text{doc2vec}(D_i) \cdot \text{doc2vec}(D_j) \tag{3}$$

The final informativeness score of a news item $I(D_i)$ was thus defined as the product of the document-level entropy and the similarity factor, discounted by a β factor (in our experiments $\beta = 1$ —given no additional information, we chose $\beta = 1$ as a mid-ground between exploration and exploitation. Further experimentation is necessary to fine-tune this parameter) that balances between exploration and exploitation (see Eq. (4)). This is the score by which news items were sorted before being presented to the expert annotator in Phase 2.

$$I(D_i) = H(D_i) \times sim(D_i, \mathbf{D})^{\beta} \tag{4}$$

Intuitively, the informativeness score can be interpreted as balancing two conflicting factors: diversity ($H(\cdot)$) versus domain consistency ($sim(\cdot)$). Using an entropy-based score encourages the model to prefer novel documents for which the uncertainty is higher. This is an indirect measure of diversity, but it is better than using explicitly the similarity measure because it directly leverages the classifier's learned hypothesis. Otherwise, documents with a very similar overall content, but that differ in a subtle way that completely changes the annotation (e.g., a negation) may not be considered. In turn, overall content is a good heuristic to detect out-of-domain documents or documents in a different language, which the rules used in the web scrapping phase were unable to filter out.

5.2. Phase 3 (Pre-annotation 5W1H)

In order to perform the third phase definition showed in Section 4.3, a model that annotated the 5W1H labels was required. The 5W1H labels consist of finding answers to the questions WHAT, WHEN, WHO, WHERE, WHY, and HOW in the news item to be annotated, as explained in Section 3. To accomplish this task, we proposed using a question answer (QA) model available at Hugging Face repository.[7] This model was built with a fine-tuned distilled version of BETO model (Canete et al., 2020) on SQuAD-es-v2.0 dataset (Rajpurkar et al., 2016) to fit in QA task.

The 5W1H examples of the previous two phases were divided into three sets (training, development, and test) with a target to adapt this model to our dataset, which is known as fine-tuning. The fine-tuning was performed through training and evaluation with the training and development set. This process was carried out using the Simple Transformers library.[8] The initial hyperparameter settings for this 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 repository.[9]

[7] https://huggingface.co/mrm8488/distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es

[8] https://simpletransformers.ai/

[9] https://gplsi.dlsi.ua.es/resources/BETO_QA_SPANISH_5W1H_fine_tuning

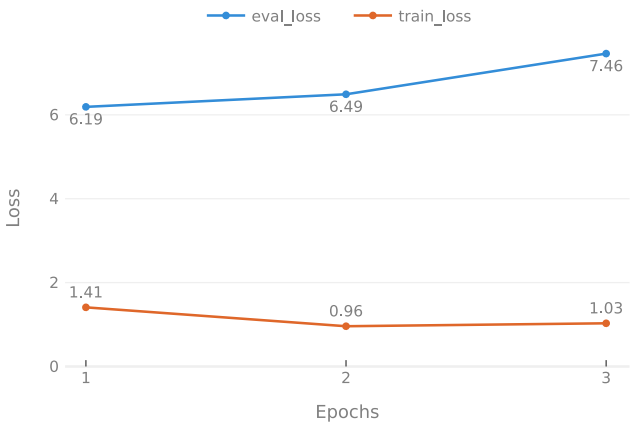


Fig. 3. Loss curve using the training and development set during training process.

The model inputs to train are the questions (5W1H), their context, and their respective answer. The model returned an answer as well as a score that represented the probability of certainty associated to the answer. Fig. 3 shows the loss curves for training and evaluation where the behavior of the model can be seen during three epochs of training.

According to the graph in Fig. 3 after the first training epoch, the loss in the training curve decreases from 1.41 to 0.96, and the loss in the evaluation curve increases from 6.19 to 6.49. This behavior remains in the third iteration, which indicates that the model is overfitting. So we selected the first iteration to annotate the 5W1H labels. At this point, we started the third phase of annotation with the 5W1H model fine-tuned.

5.2.1. Fine-tuning performance of the QA model

We compared the performance of the QA model with fine-tuning and without fine-tuning. Table 1 shows the results annotating with the QA model without fine-tuning on 5W1H and with QA with fine-tuning on 5W1H. The main metrics for the QA task – Exact Match (EM) and F1 score – are included. These metrics are used to measure ML models performance on well-known datasets such as The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). Additionally, we compute other metrics that allow us to measure similar and incorrect answers because we consider them important to assessing the quality of the pre-annotation task. The definition of each metric is:

- Exact Match (EM): the number of the exact matches of the predicted answer with the manual answers.
- Similar: the number of the partial matches of the predicted answer with the manual answers.
- Incorrect: the number of predicted answers that do not match the manual responses.
- Overall EM: the percentage of exact matches over the number of predicted examples.
- F_1: the F_1 score is the harmonic mean of the precision and recall (Grandini et al., 2020). Precision is the ratio of the number of overlapped words to the total number of words in the prediction, and recall is the ratio of the number of overlapped words to the total number of words in the ground truth (Rajpurkar et al., 2016).

Considering the QA metrics defined previously, our main objective is maximizing the number of exact matches (EM), reducing incorrect and similar matches as much as possible, as EM implies that annotators do not have to modify the pre-annotation provided by the system. As shown in the table below, the QA models with fine-tuning obtained better results on all metrics presented. The improvement is particularly

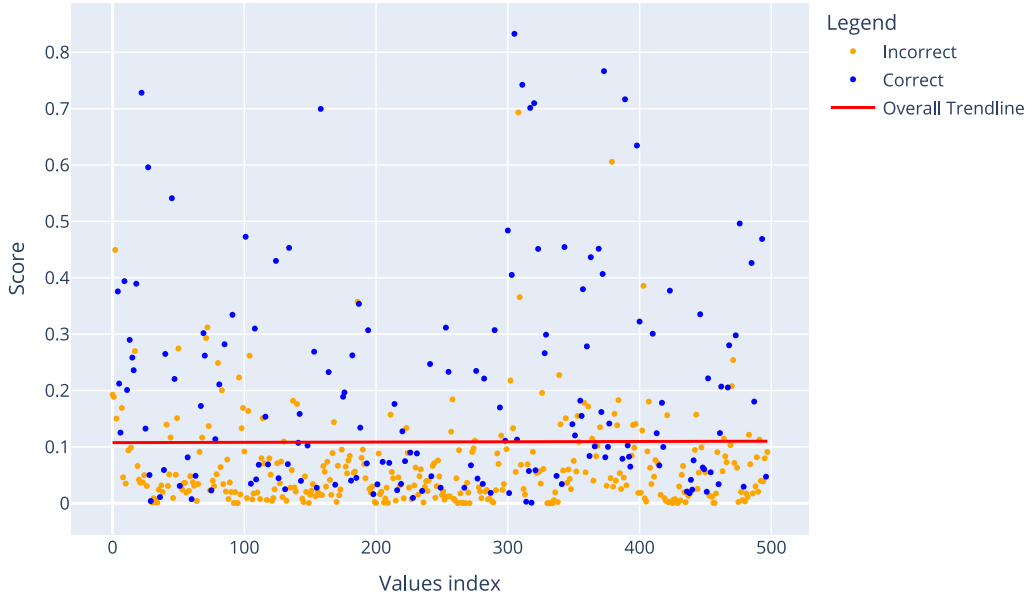


Fig. 4. Representation of 5W1H labels by QA model prediction scores, an index of each label, and the manual classification by an expert annotator of the correct and similar (blue points) and incorrect (orange points) labels.

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

Model	EM	Similar	Incorrect	Overall EM	F_1
QA without fine-tuning 5W1H	30	**396**	**141**	0.052	0.191
QA fine-tuning 5W1H after Phase 2	236	178	153	0.416	0.613
QA fine-tuning 5W1H after batch 6 in Phase 3	**263**	152	152	**0.463**	**0.641**

noteworthy in overall EM and F1 after the second fine-tuning. Consequently, these results confirm that fine-tuning is beneficial for the 5W1H label pre-annotation task.

After batch 6 of Phase 3, with 60 more news items, the 5W1H model was retrained. The new model (QA fine-tuning 5W1H after batch 6) 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 that reduces annotation times and finally assists the human annotator in this complex task.

5.2.2. Tuning the QA model threshold

Pre-annotation aims to annotate as many 5W1H elements as possible with high precision so that the annotator has to discard very few examples as incorrect, because the higher the number of corrections, the longer the delay in the annotation process. In order to reduce the error rate in the number of pre-annotated examples by the model, we automatically annotated the 5W1H labels in the 10 news items (batch 1 of Phase 3) and classified them manually as correct or incorrect. In Fig. 4, a scatter plot is used to represent each 5W1H label by an arbitrarily assigned label index (x-axis), and the score assigned by the QA model (y-axis). Finally, an ordinary least squares regression trendline was added to distinguish the correct and similar answers (blue points) from incorrect ones (orange points).

This process defined a threshold to separate the incorrect answers from correct or similar ones. In this case, the threshold selected was 0.11 because it was the closest score at all times to the ordinary least squares regression trendline, thereby separating manually classified label types. This threshold was configured in the semi-automatic annotation system with the best QA model obtained to start the annotation process in Phase 3 (QA fine-tuning 5W1H after Phase 2).

5.3. Computational prototype

The computational prototype enables the annotator to select, skip news (not interesting for the dataset), and pre-annotate news in the same interface (see Fig. 5).[10]

The user interface is the Brat tool along with the assisted system implemented which allows the annotator to discard, accept or modify the annotation proposals (see Fig. 6). The use of this interface enabled us to annotate quickly, accurately and easily.

Despite using this somewhat dated annotation software that is not as full-featured as more modern alternatives, we found that its simplicity and minimalist design is an advantage when introducing new annotators to a semi-automated workflow. Furthermore, its file-based storage model, and its open annotation syntax allowed us to integrate our semi-automated workflow without having to access or modify Brat's source code. In addition, in this phase both the compilation and the annotation tasks were carried out in the same interface, without having to switch screens and search external sites. The assisted annotation system was based on a predictive interface which, as stated by Monarch (2021), consists of items that have been pre-annotated by a machine learning model. This type of interface enables annotators to edit items, and readjust the model with the errors detected and corrected. Thanks to this navigability, the system integrated the news recommendation, which not only saved time, but also took into account the annotator's selection and, on that basis, retrained the model through active learning.

[10] The original news source of the figure is available at https://www.eldiestro.es/2021/04/mintiendo-y-manipulando-asi-pretenden-marcar-los-miserables-medios-de-comunicacion-de-espana-a-las-personas-que-decidan-no-vacunarse/

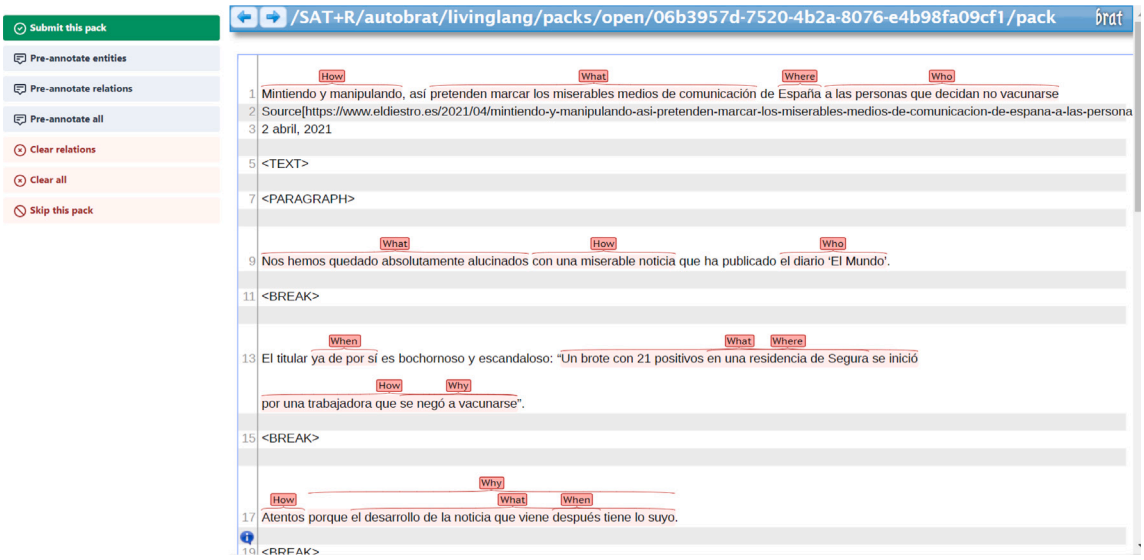


Fig. 5. Pre-annotation in Brat.

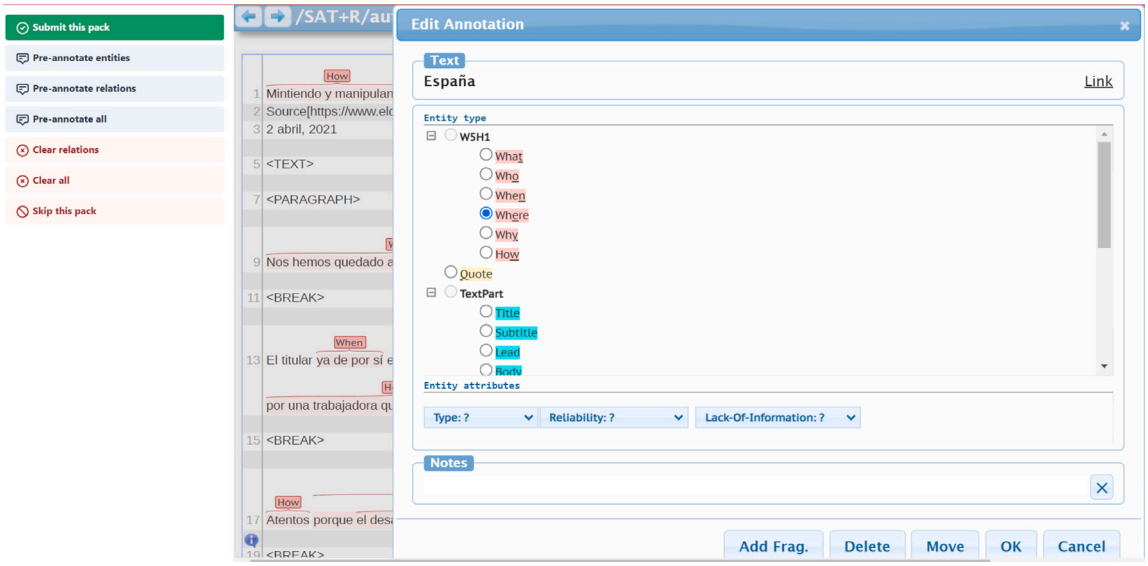


Fig. 6. Modification and selection of labels and attributes.

5.4. Generalization of the proposed methodology

To sum up, the proposed methodology comprises three phases. In Phase 1, a set of news items are manually compiled and annotated. After that, Phase 2 introduced active learning, where the human annotator interacts with the ML model that automatically selected the most informative news item to annotate. While this phase creates a more efficient process by reducing compilation costs, the annotation is still manual. Finally, Phase 3 adds human–machine interaction in the annotation process, consisting of the human reviewing and improving the automatic pre-annotation of the dataset, provided by an ML-assisted labeling system and implying a machine-human-machine loop. As stated before, the proposed methodology was implemented and applied to the reliability annotation within the disinformation framework, but it is easily applicable to a broad range of complex annotation problems, following the same phases. Some preliminary ongoing studies on other annotation schemes are being performed to confirm this fact. Regardless of the concrete entities and complex relationships that one wants to deal with in a complex annotation scheme, the entropy and informativeness formulas used in AL are not specific to the annotated entities or relationships. Likewise, the pre-annotation module can be replaced by an ML-assisted labeling to the concrete problem. Furthermore, the fact that the corpus is in the Spanish language is irrelevant to the experimental results since the machine learning models used are language-agnostic and no language-specific heuristics are applied. Hence, these results should generalize to other languages and annotation schemes albeit with different baseline F1 scores according to the complexity of the underlying learning problem.

6. Evaluation framework

In order to measure both the benefits of the methodology adopted and the quality of the resource generated, this section undertakes a set of experiments. First the features of the dataset constructed are presented. Second, two dataset quality measures are provided, both

Table 2
Numerical description of the 5W1H in RUN dataset.

5W1H	Unreliable items	Reliable items	Total items
WHAT	687	1600	2296
WHEN	117	573	690
WHERE	685	58	747
WHO	326	1525	1856
WHY	142	241	384
HOW	165	358	529
TOTAL dataset	2122	4355	6502

in terms of labeling consistency and inter-annotation agreement. Regarding the methodology, both the efficiency and effectiveness of it are measured. Finally, the limitations of the proposal are discussed.

6.1. RUN dataset description

After applying the HITL-based methodology, a Reliable and Unreliable News (RUN) Dataset in Spanish (both from Spain and Latin America) was obtained. RUN dataset consists of a set of 170 news items collected from mainstream digital media that have the traditional journalistic structure.

As indicated in Section 3, each 5W1H item is assigned a reliability value. The reliability criteria adopted for this annotation (Bonet-Jover et al., 2023) is not related to the source credibility but to the accuracy and neutrality of the semantic elements of the news item. Therefore, the dataset is balanced according to the Reliable (85) and Unreliable (85) classification, and it was created via an incremental procedure whereby 40 news items were included in Phase 1, 40 more news items were included in Phase 2 and 90 more news items in Phase 3. Due to Phase 3 being more efficient, we were able to annotate more news items in less time. The size of the dataset is limited in this initial phase of the creation, since the aim was to prove the validity of the semi-automatic building methodology proposed. Even so, given the characteristics of its creation, where human-in-the-loop strategies have been used, the size should not be a problem to demonstrate the validity of the methodology since the examples are more representative than if this same sample had been chosen randomly as it was the case when the annotation is entirely manual and no human-in-the-loop strategy is involved in the process (Monarch, 2021).

Previous research found that news mixes unreliable and reliable information, which hinders the disinformation detection task (Bonet-Jover et al., 2021). We consider that the different parts and content elements of a news item have specific reliability values that influence the global reliability value of a news item. To gauge how information is distorted in news, we need to analyze each part and component separately. This requires a balanced dataset of unreliable and reliable news to train our system. Details are presented in Table 2.

As can be seen from the figures in the Table 2, 4,355 Reliable 5W1Hs are available in the dataset compared to 2,122 Unreliable 5W1Hs. Although the news dataset is balanced between reliable and unreliable, according to the figures of this dataset, the fact that there are many more reliable labels indicates that in news intended to spread disinformation, not all labels will be unreliable, both types of information (reliable and unreliable) would be mixed in order to confuse the reader. As stated by Juez and Mackenzie (2019) "in most cases, fake news is not totally false, but rather a distorted version of something that really happened or a manipulated account of true facts". This does not mean that in a reliable news item all the labels are also reliable, some might not be, probably unintentionally, but precisely because of this, the model will be able to learn to classify into reliable or unreliable based on those data.

The aim of the present work is to increase the speed of creating the dataset without compromising accuracy. Given the time spent on the annotation task as well as the complexity and the semantic nature of our annotation guideline (which makes the agreement between annotators more complicated and subjective), the annotation methodology improved the procedure, as demonstrated in the next subsections. To validate our methodology, we do not need a huge dataset, but a quality dataset with rich and well-chosen examples that increase the accuracy.

6.2. Annotation process details

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. First annotator has a high level of expertise annotating text with this type of annotation. The second one was an experienced annotator but in other types of semantic annotations.

Since HITL techniques are techniques involving humans in the training process, we have to consider in the annotation process aspects related to human–computer interaction (HCI). Due to that, the annotation plan comprises the following elements:

- *Objectives of the annotation process:* The purpose of this annotation task is to determine the structure-type items of the inverted pyramid and the essential content given by the 5W1H items of the text and the reliability of each of these elements. All the items to be annotated are clearly defined in the annotation guideline RUN-AS (Reliable and Unreliable News Annotation Scheme).[11] The items to be annotated will be defined in the annotation tool to be selected when the annotator deems it necessary.
- *User-friendly annotation interface:* The user interface, as indicated in Section 5.3, is the Brat tool along with the assisted system implemented. This interface enables the annotator to select, skip news, and pre-annotate news in the same interface. The web tool, hosted in a server, automatically saved the work done and showed the labels with colors and symbols without having to place the cursor above for identification purposes. The interface is minimalist and functional, facilitating the onboarding of new annotators, as it is not necessary for them to deal with complex functionalities for project management or credentials. However, at the same time, Brat's annotation system is powerful enough to deal with a complex schema like ours with several different types of entities and relations.
- *Help tutorial:* A document with clear instructions on how to use the annotation tool was given to the annotators and it is available at https://gplsi.dlsi.ua.es/resources/brat_guideline.
- *Annotation sessions' planning:* Bearing in mind that the annotation process includes human intervention, either annotating from scratch as in phase 1, or assisted annotation in phase 3, and taking into account the possible fatigue of the annotators, 30 to 40 min sessions were planned in each phase. The rationale of this is trying to maximize the quality of the annotation and not the quantity of it.
- *Usability proofs:* A post-annotation usability satisfaction questionnaire to collect the opinion of the human annotators was performed. The usability questionnaire was derived from Zhang and Adams (2012). Usability testing helps us to identify problematic areas and improve the user experience. After analyzing the results, annotators somewhat agree in the usability of the tool (Lewis, 1995). All the feedback obtained from these proofs would be considered for future improvements in the annotation process. The questions used in the usability proofs are available at https://gplsi.dlsi.ua.es/resources/usability_annotation_proofs.

[11] Available at https://gplsi.dlsi.ua.es/resources/NewsReliabilityAnnotation

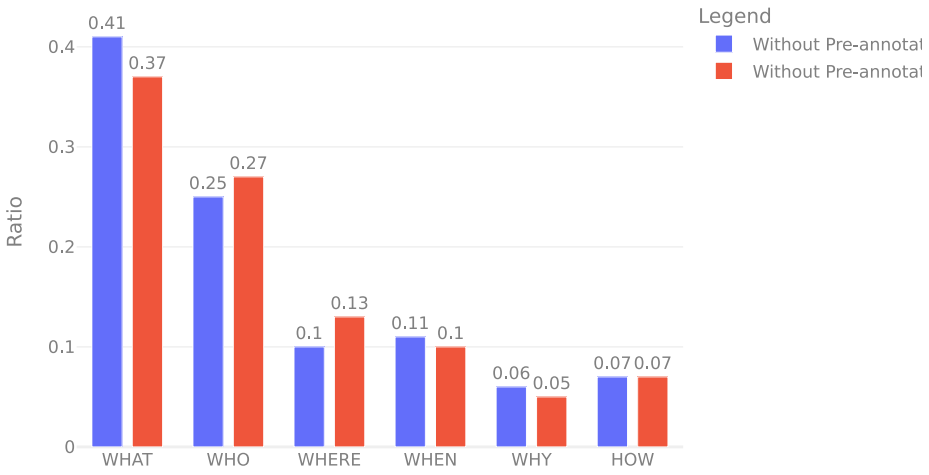


Fig. 7. Comparison of the distribution of annotations over 5W1H for batches with and without pre-annotations.

6.3. Annotation quality

To validate the generated resource built following the semi-automatic methodology and to evaluate the difference in terms of quality between batches with and without pre-annotations, we analyzed the distribution of the 5W1H annotation in Phase 1 (without pre-annotation) and Phase 3 (with pre-annotation). In addition, we assessed the inter-annotator agreement of the final RUN dataset obtained.

The distribution of 5W1H annotations is presented in Fig. 7. As can be seen, per each 5W1H label, the distribution of annotated items is quite consistent, so this slight deviation between manual and pre-annotated phases indicates the annotator performance is stable in producing annotations under such conditions.

To further assess the quality of the dataset, the inter-annotator agreement (IAA) between two expert linguistic annotators was computed following the formula used by Névéol et al. (2011): IAA = number of matches/(number of matches + number of non-matches). Other well-known IAA metrics, such as Cohen's *kappa* (Cohen, 1960) were discarded since, when large segments of text are not annotated, the degree of agreement between annotation versions may be overestimated by Kappa (Piad-Morffis et al., 2019). Therefore, this metric is not relevant for token-level annotation tasks (where many tokens are not annotated in the text), so it is unsuitable due to the characteristics of the annotation proposed here. In these cases, the agreement between two raters can be quantified using traditional information retrieval metrics (Hripcsak and Rothschild, 2005).

The data presented in Table 3 are adjusted only to the agreement in those cases where annotation by at least one of the annotators has been considered. This is intended to provide a more meaningful comparative measure.

The criteria to consider when there is an agreement or not are the following. When comparing annotations, we considered that a match occurred when the annotators (A and B) agreed on assigning the same category to a specific span of elements in the text. A slight difference in length regarding the span of the elements to be annotated is allowed, as long as one string is contained in the other one. For example: "scientists" annotated as WHO by the annotator A and "scientists specialised in biophysics" annotated as WHO by the annotator B. We considered as a non-match those cases where the annotators did not agree to use the same label for a span of elements in the text. For instance: "scientists" was annotated as WHO by the annotator A and it was annotated as WHERE by the annotator B. There would also be no match if they annotate with a 5W1H different portions of text.

Considering the high complexity of the annotation and the fact that the tokens that have not been annotated by any annotator were not included in the metric, which would favor the final results, these results are suitable. Furthermore, after analyzing the failure cases in detail, the following conclusions can be drawn. For the case of the inverted pyramid, the disagreement is mainly in the conclusions section as it is an optional section in this structure and with a high degree of subjectivity. However, IAA's result of 0.8 is adequate in this case since this type of error does not affect the performance of the system because the conclusion content is included as part of the body. As future work, it is proposed to evaluate the need for this label.

Table 3
Interannotator agreement per annotation level of RUN Dataset.

Annotation level	RUN Dataset
Inverted pyramid	0.80
5W1H	0.70
Elements of Interest	0.63
Complete annotation	0.70

On the other hand, the IAA obtained for 5W1H and EoI levels illustrates the high semantic complexity of this type of annotation. However, after calculating the IAA of the 5W1H labels when they are completely manually annotated, the resulting value obtained was 0.64. This proves that the semi-automatic annotation of the 5W1H level allows to increase the agreement with respect to the manual annotation and therefore to generate a higher quality resource.

Considering the problem of the intrinsic subjectivity of this annotation task, in future works will be addressed the automatic extraction of the most relevant content to be annotated, such as using automatic summarization, to simplify the annotation task.

6.4. Measuring efficiency of the methodology

To measure the evolution of the dataset construction with the proposed methodology we assessed the time spent on each phase, both in the compilation and in the annotation tasks, and the pre-annotation error rate for each of the 5W1H labels, to determine if all 5W1H labels are suitable to be pre-annotated.

6.4.1. Measuring time reduction

The time spent in the compilation task was calculated based on the time it took to find, read and save each news item, while the annotation task was calculated on the basis of time spent per news item's annotation. An average of the compilation and annotation time per news item (minutes per news item) is provided for each phase, as well as the total average time per news item (see Table 4). As indicated

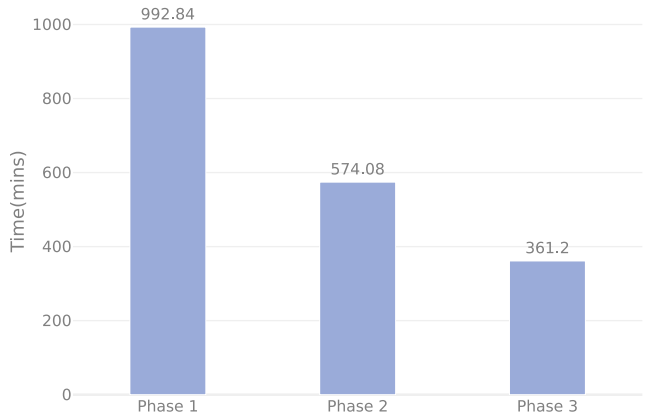


Fig. 8. Time reduction in the annotation process for each phase, considering 4 batches of 10 news items per batch with an average of 20,000 words per phase.

Table 4
Comparison of phases measured in time per news item.

Phase	Compilation (min/news item)	Annotation (min/news item)	Average Time (min/news item)
Phase 1	15	16.7	31.7
Phase 2	3	16.7	19.7
Phase 3	**1.3**	**12.3**	**13.6**

in Section 4, Phases 1 and 2 comprises 4 batches with 10 news per batch and Phase 3 were 9 batches.

Each parameter was measured per phase in order to compare the progression of the dataset construction. In Phase 1, the compilation process took an average of 15 min per news item, as the annotator had to search sources, read them to ensure they were appropriate for the dataset, save them in the computer, and add them to the interface. All of this was processed manually. As for the annotation process, news was manually annotated and, in this first stage, the average time spent in the annotation task was 16.7 minutes/news item.

In Phase 2, after evaluating the time spent on the compilation task and comparing it with Phase 1, we came to the conclusion that the assisted system saves time, since it suggests a list of news automatically and the expert only has to read the news and check whether it is suitable for the dataset. The expert does not have to spend time searching for sources and news or downloading them. In addition, repeated or already discarded news are not taken into account, which in turn saves time. However, the list of suggestions is not integrated into the Brat software, which leads to processing the list manually, without the possibility of selecting or discarding the news item and then annotating it directly on the same interface. In this stage, the average time spent in the compilation process is 3 min per news item. Hence, there is a clear difference in the compilation task between Phase 1 and Phase 2. What took two and a half hours to collect 10 news items, took 30 min with the automation of this task. Regarding the annotation task, as it is still done manually, there is no difference compared to Phase 1. The time spent in the annotation of Phase 2 was 16.7 minutes/news item.

In Phase 3, the compilation process took an average of 1.3 min per news. Since the system directly displays the proposed news item in the annotation interface, the expert only has to read it and decide whether it is valid or not, which reduces the compilation time compared to Phase 2. When it comes to the annotation task, pre-annotation reduced the annotation average time to 12.3 minutes/news item. The system navigability facilitates both tasks (compilation and annotation). Finally, the average time consumed per phase in 4 batches with 10 news per batch (40 news items/phase) is given to illustrate the time reduction (see Fig. 8). This comparison is introduced to demonstrate the benefits of the application of the proposed methodology, in comparison with fully manual annotation (Phase 1).

Table 5
Ratio of EM, Similar and Incorrect 5W1H labels of the best's performance QA model. In bold the higher values in EM and the lower values in the incorrect metric.

5W1H labels	Ratio		
	EM	Similar	Incorrect
WHAT	**0.88**	0.09	**0.04**
WHEN	**0.70**	0.12	**0.22**
WHERE	0.62	0.01	0.30
WHO	**0.84**	0.03	**0.12**
WHY	0.45	0.10	0.45
HOW	0.46	0.08	0.46

To conclude, in the annotation process there was a reduction in time of 42.17% when annotation was performed in Phase 2 compared to the fully manual annotation (992.84 mins/40 news in Phase 1 vs. 574.08 mins/40 news in Phase 2), and a final reduction of 63.61% after performing the annotation process in Phase 3 compared to the fully manual annotation (992.84 mins/40 news in Phase 1 vs. 361.25 mins/40 news in Phase 3). Therefore, annotators also became more efficient when assistance is available in the form of automatic compilation and pre-annotation.

6.4.2. Measuring pre-annotation error rate

An analysis of the errors committed by the pre-annotation model of the 5W1H for each of these labels has been carried out so that it is possible to analyze 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 type of label not to be pre-annotated, because correcting a type of label that fails may be more costly than annotating it from scratch.

Considering both the metrics and the results obtained by the QA pre-annotation model presented in Section 5.2.1, Batch 7-Phase 3 was selected to perform this analysis, because we obtained a better QA model (QA fine-tuning 5W1H after batch 6 in Phase 3). In addition, the threshold was experimentally increased from 0.11 to 0.14 to test whether the model was able to reduce the ratio of incorrectly annotated labels, which should lower annotation times.

Table 5 shows the ratio between each of the following categories for the 5W1H labels: exact match (EM), similar, or incorrect. The ratio has been calculated taking into account the number of elements in each category with respect to the total. The main objective of the QA models should be to maximize EM answers and minimize incorrect answers. It is also important that the number of similar ones should also be as low as possible, since although there is a partial success in the annotation, it implies that the annotator has to adjust the pre-annotation and therefore, it is an added annotation workload.

After analyzing the results, it is observed that the system annotates the labels WHAT, WHEN, and WHO correctly in a very high ratio on EM (0.88, 0.70, and 0.84 respectively). The results in the case of WHERE are more limited regarding EM, but it is clear that the most complex labels for the QA system to determine accurately are WHY and HOW. Although the Similar metric obtains relatively low values in general for all labels, which is very appropriate in our case, the highest Similar value is obtained for the WHEN label, which indicates that determining the bounding of a temporal expression can be sometimes confusing. Note that all labels get higher or equal ratios (in WHY and HOW labels) in the EM metric than in the incorrect one. Finally, after this analysis, we conclude that the QA model assists the annotator properly, and it is feasible to pre-annotate all 5W1H labels this way. Despite this, we would improve the QA model in the future to reduce the error rate, especially in WHERE, HOW and WHY labels.

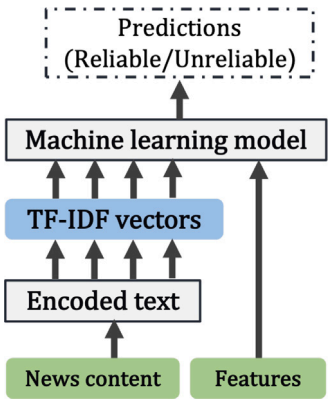


Fig. 9. Internal structure of ML baseline system.

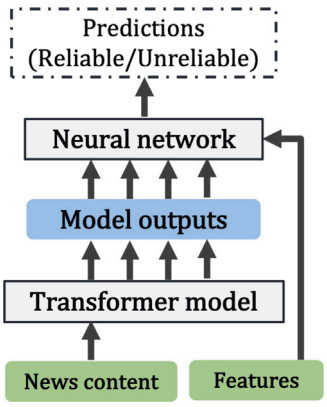


Fig. 10. Internal structure of DL baseline system.

6.5. Measuring effectiveness of the methodology

To validate the effectiveness of the methodology proposed for the construction of the dataset, the semi-automatically generated dataset is evaluated using two baseline systems to predict a random test set with 20 news items.[12] This test set was obtained before the active learning process, using the initial news pool (randomly selected) to guarantee that the test data and initial training data (Phase 1 of methodology) are from more or less the same distribution (Monarch, 2021). Figs. 9 and 10 show the internal structure of the ML and DL baseline systems, respectively.

The baseline systems were trained with the RUN dataset to predict the test set. In all cases, the baselines had as input the news content (TITLE and BODY text), and additionally, in some cases they also used as input the features obtained from the fine-grained annotation. The output of each baseline was the Reliable/Unreliable classification.

Fig. 9 represents the ML-based baseline where the news content is encoded as TF-IDF vectors and they are concatenated with the extracted features. The resulting vectors were used as input to the ML model (in this case, Logistic regression) to perform training and prediction. For this baseline, we chose TF-IDF encoding vectors and Logistic regression algorithm because, although they are a classic and essential word representation and classifier, they still performed well in similar tasks when numerical and categorical features are used, such as in the fake news detection task (Li, 2021; Jiang et al., 2021; Posadas-Durán et al., 2019; Lahby et al., 2022).

Fig. 10 represents the DL-based baseline. In this case the news content was encoded by using a transformer model. The encoded vector (model output) and the annotated features were used as input to the neural network – in this case, multilayer perceptron (MLP) – to perform training and prediction. The baseline used the classification architecture proposed by Sepúlveda-Torres et al. (2021), which combines the transformer model with external features. In this case, to design the baseline system we selected the BETO pre-trained model as the transformer model because it is a Spanish language model that obtains good performance in multiple NLP tasks (Canete et al., 2020). The BETO model is used in fine-tuning mode.

From the three annotation levels (Structure, Content, and Elements of Interest) of RUN-AS annotation, 42 numerical and categorical features were extracted. Table 6 shows the extracted features.

A simplified example of the numerical and categorical features extracted from the TITLE and LEAD of a news piece is presented next.[13]

Table 6
Overview of the 42 numerical and categorical features used.

Level	Features	
Structure (7 Categorical)	Title	Conclusion
	Subtitle	Title_Stance
	Lead	Title_Style
	Body	
Content (28 Numerical)	What	Where_Reliability_Reliable
	What_Reliability_Reliable	Where_Reliability_Unreliable
	What_Reliability_Unreliable	Where_Lack_Of_Information_Yes
	What_Main_Event	Why
	What_Lack_Of_Information_Yes	Why_Reliability_Reliable
	Who	Why_Reliability_Unreliable
	Who_Reliability_Reliable	Why_Lack_Of_Information_Yes
	Who_Reliability_Unreliable	How
	Who_Lack_Of_Information_Yes	How_Reliability_Reliable
	When	How_Reliability_Unreliable
	When_Reliability_Reliable	How_Lack_Of_Information_Yes
	When_Reliability_Unreliable	Who_Role_Subject
	When_Lack_Of_Information_Yes	Who_Role_Target
	Where	Who_Role_Both
EoI (7 Numerical)	Quote	Quote_Author_Stance_Agree
	Quote_Author_Stance_Disagree	Quote_Author_Stance_Unknown
	Key_Expression	Orthotypography
	Figure	

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

The same feature types were generated from the other parts of the inverted pyramid structure of the document. Each feature indicates the number of 5W1H components with a specific label and reliability attribute that appear in each part of the news. For example, `LEAD_WHAT_Reliable: 2` indicates that the LEAD contains two

[12] Available at https://gplsi.dlsi.ua.es/resources/NewsReliabilityAnnotation

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

Table 7
Performance of ML and DL models for predicting test set with features and without them.

Model	Without features		With features	
	F_1	Acc	F_1	Acc
Logistic Regression (TF-IDF encoded)	0.733	0.75	**0.949**	**0.95**
BETO	0.84	0.85	**0.89**	**0.9**

`WHAT` items annotated with a `Reliable` value. The models were trained to predict the overall document reliability label based on these numerical and categorical features.

Table 7 shows the results of the baselines in terms of the metric F_1 and accuracy to predict the test set. In order to evaluate the contribution provided by the annotation used, two experiments with each baseline were performed. The first one used only the news content (Without features column), and the second one concatenated the extracted features of the RUN dataset (With features column). To replicate the results shown in Table 7 the following GitHub repositories can be used ML-based baseline[14] and DL-based baseline.[15]

As can be seen in the table, both baselines when using RUN-AS features outperformed the baseline without features. The Logistic Regression model performed better than BETO when using the features. Furthermore, BETO model performed better than Logistic Regression when not using RUN-AS features, showing the power of pre-trained models when they were not using specific features. This result corroborates not only the benefits of this fine-grained annotation (RUN-AS scheme) to reliability detection but also the feasibility of the semi-automatic dataset generated by applying the HITL-based methodology. The dataset construction time was reduced by around 64% without compromising performance (95% F1 and Acc). To support this statement, a comparison with a previous research (Bonet-Jover et al., 2023), in which the dataset was built entirely manually, is done as a baseline for performance since that previous dataset was costlier and less efficient to build than the one proposed here. This is not a direct comparison since the previous dataset is different in size. It presented only 80 Reliable and Unreliable news items. However, the same annotation guideline was used. The results obtained for Logistic Regression and BETO were 0.88 and 0.85, respectively. As it can be observed, results slightly increased in the current proposal since the dataset is composed of more examples after applying the methodology that enables a more efficient construction of the dataset.

6.6. Discussion and limitations of the proposal

A limitation of our proposal is related to using active learning to select the best news examples. This is because the active learning models inherently learn about the annotator's preferences. As the model's training continues, the model learns the annotator's biases through this interactive process, and this can lead to a cycle of positive reinforcement whereby the model proposes only those documents that the annotator considers valuable or important.

However, this may be mitigated in two ways. Firstly, by always introducing an element of randomness, and not necessarily choosing only the K most informative news items that the model proposes, but including also random documents. Secondly, by having a variety of annotators and not one annotator alone to train the AL model, combining what they are annotating so that the biases are somewhat counteracted.

This discussion of bias occurs for models in general, with the addition that since the model is being trained interactively and its output is used to guide the annotator, the positive reinforcement cycles of biases may be shorter.

Furthermore, in this phase of the work, a simple AL model based on logistic regression was used, and although it allowed us to obtain positive results, in the future we will consider using other ML or DL approaches and even other HITL strategies. Likewise, the pre-annotation had a considerable error rate for more complex labels such as WHY and HOW, whose pre-annotation should be studied in depth in order to be improved.

Despite the dataset's limited size at present, namely comprising a set of 170 news items in Spanish, one of the basic principles in human-in-the-loop techniques – and more specifically in active learning (see Section 2.2) – is precisely to reach the target performance for a machine learning model faster, since the best examples are automatically obtained, and, indeed, a high performance was obtained with this corpus. After analyzing the experimentation results, the size was proved to be sufficient for the purpose of this work, which was to determine the validity of the proposed methodology, in part thanks to its form of creation with the application of the Human in the loop paradigm. However, this is currently a preliminary study and although the results are promising, the dataset would need to be extended to determine how it responds in the case of a large-scale dataset. Furthermore, in case the results hold with a larger dataset, the next step would be to analyze how to apply them to more complex tasks such as veracity detection. Lastly, regarding the computational prototype, a different annotation software could be considered in the future, given the fact that BRAT is appropriate but limited in some aspects compared to more modern alternatives.

7. Conclusions and future work

The main novelty of this proposal is the design and implementation of a human-in-the-loop based methodology for semi-automatic annotation of semantically complex datasets. Firstly, the methodology applied Active Learning (AL), a well-known strategy, to determine the most suitable news to be annotated. Specifically, the diversity sampling strategy was performed. Secondly, a human–machine interaction procedure was implemented consisting of the human reviewing and improving the automatic pre-annotation provided by the ML-assisted labeling system. This was carried out to: (i) enhance the reliability detection classification by improving the training dataset; and, (ii) re-train the ML-assisted labeling system (a QA system in our case) with the new reviewed examples.

The application of the methodology results in an improvement of the annotation task that is derived from two independent factors: intelligently sorting which news to annotate and providing pre-annotated suggestions with a high degree of certainty.

The methodology is implemented in the disinformation framework, specifically in the news content's reliability in Spanish language, producing the RUN dataset. The building of this dataset using the methodology, although limited in size at its current stage, constitutes a proof of concept of how the application of this methodology in building much more bigger datasets could significantly reduce the cost of dataset creation. Furthermore, this dataset is also a novel contribution since state-of-the-art disinformation datasets are annotated with a unique veracity value for the whole news item, whereas in our annotation proposal, essential content of each news item is identified and classified as reliable or unreliable. This is more suitable for applying explainable AI in future work, providing the evidences of disinformation for each news item.

The performance of the methodology is evaluated in terms of a balance between time-effort consumption and accuracy achievement. The annotation procedure time was reduced by 63.61% with respect to the fully manual annotation, and the inter-annotator agreement was increased. Regarding the performance of the dataset in the reliable-unreliable classification task, the baseline ML-model trained with the

[14] https://gplsi.dlsi.ua.es/resources/Logistic_Regression_RUN_Dataset
[15] https://gplsi.dlsi.ua.es/resources/BETO_RUN_AS

semiautomatic annotated dataset obtained an accuracy of 95%, which indicates high performance of the resource for the set task.

The following ongoing and future open challenges are presented:

- The initial version of the dataset, created to validate the feasibility of the proposal is limited in size. Ongoing research consists of applying the methodology to increase the dataset size and evaluate it in future models to detect disinformation.
- A further improvement of the dataset generation efficiency by reducing the size of the news items to the essential content and annotating this summarized content. Apart from reducing the dataset generation time, this is likely to improve the inter-annotator agreement and the accuracy of the trained model as more examples would be provided.
- Further research can apply the methodology to other languages and other complex annotation problems in different domains, such as hate and violent speech, which is also part of our research (Botella et al., 2023), confirming that the proposed methodology is generalizable. Furthermore, different implementations of Phase 2 and 3 could be applied.
- Further work can experiment with other Human-in-the-Loop techniques, such as Interactive Machine Learning, whereby the human not only gives feedback over the data but also participates in the model's features selection, with the aim of improving the efficiency of the final ML model.

CRediT authorship contribution statement

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

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

Codes used and the dataset obtained are shared in the GitHub repositories.

Acknowledgments

This research work is funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union" or by the "European Union NextGenerationEU/PRTR" through the project TRIVIAL: Technological Resources for Intelligent VIral AnaLysis through NLP (PID2021-122263OB-C22) and the project SOCIALTRUST: Assessing trustworthiness in digital media (PDC2022-133146-C22). It is 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

Abacha, A.B., Dinh, D., Mrabet, Y., 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In: Conference on Artificial Intelligence in Medicine in Europe. Springer, pp. 238–242.

Alex, B., Grover, C., Shen, R., Kabadjov, M., 2010. Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs. In: Proceedings of the Fourth Linguistic Annotation Workshop. pp. 29–37.

Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T., 2014. Power to the people: The role of humans in interactive machine learning. AI Mag. 35 (4), 105–120.

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

Benedikt, L., Joshi, C., Nolan, L., Henstra-Hill, R., Shaw, L., Hook, S., 2020. 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, New York, NY, USA, pp. 488–497. http://dx.doi.org/10.1145/3377325.3377489.

Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., Cumbreras, M.Á.G., 2021. Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. Expert Syst. Appl. 169, 114340. http://dx.doi.org/10.1016/j.eswa.2020.114340.

Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., Barco, P.M., 2023. Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation. Procesamiento del Lenguaje Natural 70, 15–26.

Botella, B., Sepúlveda-Torres, R., Martínez-Barco, P., Saquete Boró, E., et al., 2023. Violencia Identificada en el Lenguaje (VIL). Creación de recurso para mensajes violentos. Procesamiento del Lenguaje Natural 70, 187–198.

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 71, 102062. http://dx.doi.org/10.1016/j.media.2021.102062, URL: https://www.sciencedirect.com/science/article/pii/S1361841521001080.

Canete, J., Chaperon, G., Fuentes, R., Pérez, J., 2020. Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR, Vol. 2020.

Cañizares-Díaz, H., Piad-Morffis, A., Estevez-Velarde, S., Gutiérrez, Y., Cruz, Y.A., Montoyo, A., Muñoz, R., 2021. 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, pp. 216–225.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Measur. 20 (1), 37.

Daniel, A.M., 2021. Human-in-the-loop disinformation detection: Stance, sentiment, or something else? arXiv abs/2111.05139.

Demartini, G., Mizzaro, S., Spina, D., 2020. Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. IEEE Data Eng. Bull. 43 (3), 65–74.

Dhoju, S., Main Uddin Rony, M., Ashad Kabir, M., Hassan, N., 2019. Differences in health news from reliable and unreliable media. In: Companion Proceedings of the 2019 World Wide Web Conference. pp. 981–987.

Evrard, M., Uro, R., Hervé, N., Mazoyer, B., 2020. French tweet corpus for automatic stance detection. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6317–6322.

Fails, J.A., Olsen, D.R., 2003. Interactive machine learning. In: Proceedings of the 8th International Conference on Intelligent User Interfaces. IUI '03, Association for Computing Machinery, New York, NY, USA, pp. 39–45. http://dx.doi.org/10.1145/604045.604056.

Fanton, M., Bonaldi, H., Tekiroglu, S.S., Guerini, M., 2021. Human-in-the-loop for data collection: A multi-target counter narrative dataset to fight online hate speech. arXiv preprint arXiv:2107.08720.

Färber, M., Burkard, V., Jatowt, A., Lim, S., 2020. A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 3007–3014.

Feller, D.J., Zucker, J., Srikishan, B., Martinez, R., Evans, H., Yin, M.T., Gordon, P., Elhadad, N., et al., 2018. Towards the inference of social and behavioral determinants of sexual health: Development of a gold-standard corpus with semi-supervised learning. In: AMIA Annual Symposium Proceedings, Vol. 2018. American Medical Informatics Association, p. 422.

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: An overview. arXiv preprint arXiv:2008.05756.

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

Hripcsak, G., Rothschild, A.S., 2005. Agreement, the f-measure, and reliability in information retrieval. J. Am. Med. Inf. Assoc. 12 (3), 296–298.

Hsueh, P.-Y., Melville, P., Sindhwani, V., 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. pp. 27–35.

Ireton, C., Posetti, J., 2018. Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training. Unesco Publishing.

Jiang, T., Li, J.P., Haq, A.U., Saboor, A., Ali, A., 2021. A novel stacking approach for accurate detection of fake news. IEEE Access 9, 22626–22639.

Juez, L.A., Mackenzie, J.L., 2019. Emotion, lies, and "bullshit" in journalistic discourse. Ibérica (38), 17–50.

Jung, W., Jazizadeh, F., 2019. Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. Appl. Energy 239, 1471–1508. http://dx.doi.org/10.1016/j.apenergy.2019.01.070.

Kholghi, M., Sitbon, L., Zuccon, G., Nguyen, A., 2016. Active learning: A step towards automating medical concept extraction. J. Am. Med. Inf. Assoc. 23 (2), 289–296.

Kholghi, M., Sitbon, L., Zuccon, G., Nguyen, A., 2017. Active learning reduces annotation time for clinical concept extraction. Int. J. Med. Inf. 106, 25–31. http://dx.doi.org/10.1016/j.ijmedinf.2017.08.001, URL: https://www.sciencedirect.com/science/article/pii/S1386505617302009.

Lahby, M., Aqil, S., Yafooz, W.M., Abakarim, Y., 2022. Online fake news detection using machine learning techniques: A systematic mapping study. In: Combating Fake News with Computational Intelligence Techniques. Springer, pp. 3–37.

Lewis, J.R., 1995. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. Int. J. Hum.-Comput. Interact. 7 (1), 57–78. http://dx.doi.org/10.1080/10447319509526110.

Li, K., 2021. HAHA at FakeDeS 2021: A fake news detection method based on TF-IDF and ensemble machine learning. In: IberLEF@ SEPLN. pp. 630–638.

Mitra, T., Gilbert, E., 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 9. pp. 258–267.

Monarch, R.M., 2021. Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI. Simon and Schuster.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á., 2022. Human-in-the-loop machine learning: A state of the art. Artif. Intell. Rev. 1–50.

Névéol, A., Doğan, R.I., Lu, Z., 2011. Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction. J. Biomed. Inf. 44 (2), 310–318.

Okoro, E., Abara, B., Umagba, A., Ajonye, A., Isa, Z., 2018. A hybrid approach to fake news detection on social media. Nigerian J. Technol. 37 (2), 454–462.

Olsson, F., 2009. A Literature Survey of Active Machine Learning in the Context of Natural Language Processing. Technical Report, Swedish Institute of Computer Science.

Pérez-Rosas, V., Mihalcea, R., 2015. Experiments in open domain deception detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1120–1125.

Piad-Morffis, A., Gutiérrez, Y., Muñoz, R., 2019. A corpus to support ehealth knowledge discovery technologies. J. Biomed. Inf. 94, 103172.

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

Rahman, M.M., Kutlu, M., Elsayed, T., Lease, M., 2020. Efficient test collection construction via active learning. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. pp. 177–184.

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P., 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Ramos, G., Meek, C., Simard, P., Suh, J., Ghorashi, S., 2020. Interactive machine teaching: A human-centered approach to building machine-learned models. Human–Comput. Interact. 35 (5–6), 413–451.

Salem, F.K.A., Al Feel, R., Elbassuoni, S., Jaber, M., Farah, M., 2019. Fa-kes: A fake news dataset around the syrian war. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13. pp. 573–582.

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

Sepúlveda-Torres, R., Saquete Boró, E., et al., 2021. GPLSI team at CheckThat! 2021: Fine-tuning BETO and RoBERTa. CEUR.

Settles, B., Craven, M., Ray, S., 2007. Multiple-instance active learning. Adv. Neural Inf. Process. Syst. 20.

Shahi, G.K., Nandini, D., 2020. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. arXiv preprint arXiv:2006.11343.

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T.H., Ding, K., Karami, M., Liu, H., 2020. Combating disinformation in a social media age. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 10 (6), e1385.

Silva, R.M., Santos, R.L., Almeida, T.A., Pardo, T.A., 2020. Towards automatically filtering fake news in portuguese. Expert Syst. Appl. 146, 113199.

Simard, P.Y., Amershi, S., Chickering, D.M., Pelton, A.E., Ghorashi, S., Meek, C., Ramos, G., Suh, J., Verwey, J., Wang, M., et al., 2017. Machine teaching: A new paradigm for building machine learning systems. arXiv preprint arXiv:1707.06742.

Spina, D., Peetz, M.-H., de Rijke, M., 2015. Active learning for entity filtering in microblog streams. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15, Association for Computing Machinery, New York, NY, USA, pp. 975–978. http://dx.doi.org/10.1145/2766462.2767839.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J., 2012. brat: A web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Avignon, France, pp. 102–107, URL: https://www.aclweb.org/anthology/E12-2021.

Tchoua, R., Ajith, A., Hong, Z., Ward, L., Chard, K., Audus, D., Patel, S., de Pablo, J., Foster, I., 2019. Active learning yields better training data for scientific named entity recognition. In: 2019 15th International Conference on EScience. EScience, IEEE, pp. 126–135.

Thomson, E.A., White, P.R., Kitley, P., 2008. "Objectivity" and "hard news" reporting across cultures: Comparing the news report in English, French, Japanese and Indonesian journalism. Journalism Stud. 9 (2), 212–228.

Tomanek, K., Wermter, J., Hahn, U., 2007. 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, pp. 486–495.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Vlachos, A., Riedel, S., 2014. Fact checking: Task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. pp. 18–22.

Voorhees, E.M., 2018. On building fair and reusable test collections using bandit techniques. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 407–416.

Vu, H.-T., Gallinari, P., 2006. A machine learning based approach to evaluating retrieval systems. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. pp. 399–406.

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

Wang, W., Zhao, D., Zou, L., Wang, D., Zheng, W., 2010. Extracting 5W1H event semantic elements from Chinese online news. In: International Conference on Web-Age Information Management. Springer, pp. 644–655.

Wardle, C., et al., 2018. Information Disorder: The Essential Glossary. Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School, Harvard, MA.

Wondimu, N.A., Buche, C., Visser, U., 2022. Interactive machine learning: A state of the art review. arXiv preprint arXiv:2207.06196.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L., 2022. A survey of human-in-the-loop for machine learning. Future Gener. Comput. Syst..

Zhang, T., Adams, J., 2012. Evaluation of a geospatial annotation tool for unmanned vehicle specialist interface. Intl. J. Hum.–Comput. Interact. 28, 361–372. http://dx.doi.org/10.1080/10447318.2011.590122.

Zhang, H., Chen, X., Ma, S., 2019. Dynamic news recommendation with hierarchical attention network. In: 2019 IEEE International Conference on Data Mining. ICDM, IEEE, pp. 1456–1461.

Zhang, H., Liu, H., 2016. Visualizing structural "inverted pyramids" in English news discourse across levels. Text Talk 36 (1), 89–110.