IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 June 2023, accepted 12 July 2023, date of publication 14 July 2023, date of current version 19 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3295781

RESEARCH ARTICLE

Detecting Misleading Headlines Through the Automatic Recognition of Contradiction in Spanish

ROBIERT SEPÚLVEDA-TORRES, ALBA BONET-JOVER, AND ESTELA SAQUETE
Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain

Corresponding author: Estela Saquete (stela@dlsi.ua.es)

This research work is funded by Ministry of Science and Innovation of Spain, and State Research Agency (MCIN/AEI/10.13039/501100011033) and, as appropriate, by "European Regional Development Fund, 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 Natural Language Processing (PID2021-122263OB-C22) and the project SOCIALTRUST: Assessing trustworthiness in digital media (PDC2022-133146-C22). Also funded by Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/21), and the grant ACIF/2020/177.

ABSTRACT Misleading headlines are part of the disinformation problem. Headlines should give a concise summary of the news story helping the reader to decide whether to read the body text of the article, which is why headline accuracy is a crucial element of a news story. This work focuses on detecting misleading headlines through the automatic identification of contradiction between the headline and body text of a news item. When the contradiction is detected, the reader is alerted to the lack of precision or trustworthiness of the headline in relation to the body text. To facilitate the automatic detection of misleading headlines, a new Spanish dataset is created (ES_Headline_Contradiction) for the purpose of identifying contradictory information between a headline and its body text. This dataset annotates the semantic relationship between headlines and body text by categorising the relation between texts as *compatible*, *contradictory* and *unrelated*. Furthermore, another novel aspect of this dataset is that it distinguishes between different types of contradictions, thereby enabling a more fine-grain identification of them. The dataset was built via a novel semi-automatic methodology, which resulted in a more cost-efficient development process. The results of the experiments show that pre-trained language models can be fine-tuned with this dataset, producing very encouraging results for detecting incongruency or non-relation between headline and body text.

INDEX TERMS Annotation guideline, contradiction detection, dataset annotation, deep learning techniques, disinformation detection, human language technologies, and natural language processing.

I. INTRODUCTION

Digitally accessing news is common practice in present-day society. Both for digital and traditional newspapers, the headline is an essential part of a news story, as it summarises the content and gives the reader a preview of the article [1], [2]. Headlines aim to draw attention to the news quickly, briefly, and effectively [2]. They support the veracity of the whole news item without compromising precision or being misleading [3]. Through headlines, readers choose whether or not to read a news item in its entirety [4]. Unfortunately, when headlines are more focused on hooking the reader than on the accuracy of the ideas put forward, this often results in the creation of misleading headlines. These represent a recognised problem within the general disinformation phenomenon that requires detection [5].

Recent studies suggest that the first impression obtained when reading a headline influences the conclusions reached after reading the full article [2] and can even persist even when the information is contradictory between the headline

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

and the content of the news [5], [6]. Added to this is the common behavior of many users who share news even without reading the full content of the news and without verifying its veracity [7]. In traditional print media, the reader has the headline and the content of the news visible together, an aspect that has changed radically in digital media, where the reader has to click on a link to check the content of the news. Therefore, there is a certain tendency to infer the content of the information by means of the headline without reading the news in depth. Thus, the risk of not detecting contradictions increases considerably [3]. The aforementioned conclusions in the cited literature make inaccurate or false headlines a serious problem that is negatively impacting information society.

Given the existing classifications of distorting headlines in the state of the art, we decided to group them into two categories —clickbait and misleading headlines— in line with how the headline detection task is tackled. Some overlap between them is also possible. Next, a brief explanation of their characteristics follows:

- **Clickbait headlines**: Clickbait refers to content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page so as to monetize the "views" through advertising revenue (the more clicks, the more money earned). This type of headline is often ambiguous and exhibits a particular writing style to directly exploit human curiosity by, for instance, using exclamatory or interrogative headlines that urge audiences to click on the link to discover the missing information [8]. Typically, clickbait headlines are spread on social media in the form of short teaser messages that may read like the following cited examples:
  - **Headline:** "La nueva vida de Iker Casillas tras su divorcio de Sara Carbonero: esto es lo que se ha comprado" *(The new life of Iker Casillas after his divorce from Sara Carbonero: this is what he bought himself)*[1]
  - **Headline:** "El fantástico hilo de Twitter que habla de "LA NUEVA SEPA" y que debería ser leído por todo aterrado tragacionista"*(The fantastic Twitter thread that talks about "THE NEW STRAIN" and that should be read by all the terrified and gullible)*[2]

  Existing methods for automatically detecting clickbait headlines exclusively focus on the headline (its writing style or structure) rather than considering the content of the news itself, so evidence is not required [9], [10]. Furthermore, this task is usually treated as a classification problem (clickbait/non-clickbait).
- **Misleading headlines**: These headlines significantly misrepresent the findings reported in the news article, by exaggerating or distorting the facts described in the body text [6]. The reader can only discover the inconsistencies after reading the entire article [8].

  Despite the literature sometimes referring to these headlines as *incongruent headlines* [11], the distinction between both definitions is not highly marked. Therefore, our work considers incongruent headlines to be included within the *misleading headlines* concept given that the term is more comprehensive. Moreover, we are not only addressing inaccurate or contradictory headlines but also unrelated ones *(unrelated headline)* (whereby the headline does not correspond to the topic of the content) because these types of headlines also mislead the reader [12]. The unrelated category is also common when misleading headlines are addressed from the perspective of stance detection using datasets such as Emergent and Fake News Challenge (FNC-1).

  In a misleading headline, some important nuances that are part of the news body text are missing in the headline, causing the reader to come to the wrong conclusion. In contrast to clickbait headlines, the language used does not necessarily incite the reader to click on it, but it is designed to trigger emotion or excitement and as indicated in [11], the analysis of this type of headline is beyond the stylistic features of the headline. Misleading headline detection implies carrying out a semantic analysis of the relationship between the headline and the body text, unlike clickbait headlines, which exhibit well-defined structures so that solutions at the syntactic level can be effective. An example of a misleading headline is shown below:
  - **Headline:** "Selena Gomez se atreve y les enseña: la foto más Íntima" *(Selena Gomez dares and teaches them: the most intimate photo)*[3]

    Evidence within the article: "Un dı́a más, la cantante está jugando con todos sus seguidores. La artista continúa llenando estratégicamente su Instagram con fotos de ella cuando era pequeña. Curiosamente, sigue la misma metodologı́a todos los dı́as. Publica hasta tres fotografı́as. En el lado derecho vemos algunas imágenes con un mensaje y nada más. En el centro aparece una fotografı́a de Selena Gomez cuando era niña con un mensaje y en el lado izquierdo más fotografı́as actuales en blanco y negro con otro mensaje claro. ¿Qué está haciendo? ¿Alguien ya lo sabe?". *(On another day, the singer is playing with all her followers. The artist continues to strategically fill her Instagram with photos of herself as a child. Curiously, she follows the same strategy every day. She posts up to three pictures. On the right side we see some images with a message and nothing else. In the middle is a picture of Selena Gomez as a child with a message,*

[1] https://www.lne.es/vida-y-estilo/gente/2021/04/14/nueva-vida-iker-casillas-divorcio-48374650.html (accessed online 1 February, 2023)

[2] https://www.eldiestro.es/2021/04/el-fantastico-hilo-de-twitter-que-habla-de-la-nueva-sepa-y-que-deberia-ser-leido-por-todo-aterrado-tragacionista/ (accessed online 1 February, 2023)

[3] https://newsbeezer.com/chile/selena-gomez-se-atreve-y-les-ensena-la-foto-mas-intima/ (accessed 1 February 2023)

> *and on the left side there are more current black and white pictures with another clear message. What is she doing, does anyone know yet?)*

As can be seen in the example presented, the headline leads one to expect other types of photos, when the truth is that they are simply childhood photos.

Given that the treatment of misleading headlines is complex and therefore best dealt with in parts [11], as detailed in the literature review section, this paper will deal with the problem from a semantic perspective, and more specifically from the detection of contradictions. Accordingly, this research focuses on developing an effective contradiction resource with the different types of contradictions between the headline and the body text annotated. This resource will enable the detection of contradictions and will support the task of misleading headline identification.

According to [13], in a strict logical definition of contradiction, sentences A and B are contradictory if there is no possible world in which A and B are both true. However, this strict definition is relaxed in a human environment, so contradiction occurs when two sentences are extremely unlikely to be true simultaneously. To deal with the problem of contradiction detection from a computational perspective, Artificial Intelligence (AI) and Natural Language Processing (NLP) are required. At present, AI cannot learn by itself and needs to be nourished by examples created by humans [14]. Indeed, when a problem is approached from the AI perspective, either with Machine Learning (ML) or Deep Learning (DL) techniques, millions of instances of human feedback are required to get the annotated datasets that will be used to train and evaluate the systems that will be in charge of solving the problem [15]. An efficient dataset would be one that can be created as quickly and inexpensively as possible, without deteriorating performance. Moreover, this proposal has an added challenge regarding languages other than English, where there exists a scarcity of data.

Given this context, the main objective of the work presented consists of addressing the task of detecting misleading headlines by using fine-grained contradiction detection (initially addressed in [16]). The automatic detection of contradictions would help to identify unreliable information as finding contradictions between two pieces of information related to the same fact would be an indication that at least one of them contains demonstrable elements of falsehood or that the information is wrong. This would cast doubts on the reliability of the content in question, and would therefore highly benefit the fight against disinformation and its viralization. Thus, the main contribution of this work is the design of a novel methodology for the semi-automatic construction of a dataset that enables us to efficiently and effectively annotate the most important types of contradictions that exist between texts. This fine-grained annotation contributes towards explainable AI, since, as in the case of other problems, not only is relevant to detect that a contradiction exits but also the rationale for why there is a contradiction. In this way, critical thinking of technology users is triggered. Furthermore, motivated by the scarcity of non-English language data, the methodology makes a worthwhile contribution to generating Spanish contradiction resources that are sourced directly from original Spanish content.

The main application of this proposal is a step that makes headway in addressing the demand for misleading information detection. Since disinformation is considered one of the main threats to democratic countries[4], any AI technological effort that contributes to the fight against disinformation is key to maintaining healthy democracies. In addition, the problem of disinformation is likely to be exacerbated by the open use of AI tools such as ChatGPT[5]. Researchers predict that generative technology could make disinformation cheaper and easier to produce for an even larger number of conspiracy theorists and spreaders of disinformation [17]. Considering this threat, the application of the technology proposed here could involve generating an alert of contradictory information for users, thereby preventing disinformation from being distributed and mitigating the damage that the viralization of disinformation generates in society. Although content providers may not be interested in this technology as they often gain from misleading headlines, the general public could benefit by integrating it via a browser plugin, or through a bot in social networks like WhatsApp[6]. In this way, the technology would be simple and free to use for the general public, whereby given the url of a news item provided, the user could be given a report of possible contradictions between the headline and body text. It could even be used to compare two news items from different media about the same fact.

The paper is structured as follows: Section II presents the literature review regarding misleading headline detection, contradiction detection literature, and dataset construction methodologies, Section III presents the methodology to create the dataset, Section IV describes the obtained dataset and the annotation validation, Section V describes the experiments carried out and the models used, Section VI reports and discusses the results of the proposed experiments on the dataset created, and Section VII presents conclusions and outlines the main direction for future work.

II. LITERATURE REVIEW

To identify misleading headlines, this research applies contradiction detection and it is focused on the methodology to create a suitable dataset from a semi-automatic procedure. Therefore, the background section is organized as follows. Firstly, research regarding misleading headlines detection is analysed, secondly, the contradiction detection literature is reviewed, and finally the state-of-the-art dataset construction methodologies are presented.

[4] https://digital-strategy.ec.europa.eu/fr/node/1503/printable/pdf
[5] https://chat.openai.com/
[6] https://godinabox.co/

A. MISLEADING HEADLINES DETECTION APPROACHES

According to [11], the misleading or incongruent headlines should be considered a problem beyond clickbait detection. Besides, the problem of analysing the relationship between headline-body text is best approached in parts. There are approaches based on extracting key quotes [18] or claims [19], [20]. In some research an artificial headline is created from the body text using natural language generation and compared with the original headline [21]. Argument analysis is also applied to this task, detecting headlines that represent an argument that is not supported in the body text [22].

Most solutions to this task are approached from a stance detection perspective. Stance detection can be defined as the task of identifying the perspective of an author or text against a given target in the form of one topic, claim, headline or even a personality [23], [24]. For misleading headlines, it involves classifying the stance of the article's body text with respect to the claim made in the headline into one of the following four classes: a) *agrees*—agreement between body text and headline; b) *disagrees*—disagreement between body text and headline; c) *discusses*—same topic discussed in body text and headline, but no position taken; and, d) *unrelated*—different topic discussed in body text and headline. In this sense, considerable research uses the stance datasets Emergent [25] or its extended version FNC-1 [26] to create misleading headline detection approaches. Some research using these datasets are [27], [28], and [29]. Although it is a methodology widely used in the treatment of misleading headlines, as [11] indicated, determining the stance between headline and body text may not carry enough weight to determine incongruency between the two textual elements. Different works tackled misleading headlines by addressing congruency between headline and body, extracting features based on the congruence [8] or annotating million-scale pairs of news headline and body text with the incongruity label to train different neural networks [30].

Finally, there are existing works focused on extracting semantic relations between texts, similar to tasks like: recognizing textual entailment [31], contrast and contradiction detection [32]. In our research, we consider that contradiction detection is more suitable than stance detection for detecting misleading headlines. The rationale being that incongruity by definition implies incompatibility and therefore contradictory information.

Following the research line of semantic relations, and taking into account that contradiction detection can help in the detection of misleading headlines, the next section presents a review of contradiction detection methods and of the main existing resources in different languages for contradiction detection.

B. CONTRADICTION DETECTION RESOURCES

Since most of the approaches for contradiction detection are evaluated in a Natural Language Inference (NLI) framework, different resources regarding this task are presented. Regarding NLI resources, currently, the large annotated datasets for contradiction detection are mainly available in the English language [33], such as SNLI [34] and MultiNLI [35]. However, there does exist a cross-lingual dataset XNLI [36]. These three datasets have enabled the training of complex deep learning systems, which require very large corpora for successful results.

To the authors' knowledge, there are few studies that address contradiction detection in non-English languages, apart from those done by [33], [37], and [38]. Machine translation of the SNLI dataset from English to German was carried out by [33]. These authors built a model using the German version of SNLI and the prediction results were very similar to the same model trained on the original English version of SNLI. Takabatake et al. [37] created a large-scale database of pairs of contradictory events in Japanese. This database was used to generate consistent statements for a dialogue system. Rahimi and Shamsfard [38] performed machine translation into the Persian language of a subset of examples from the SNLI and MNLI corpora. This Persian language dataset was used to create a contradiction detection system.

From the multilingual perspective, the cross-lingual XNLI dataset was created in [36]. The dataset is divided into three parts: training, development and test. The training set was developed in English, and the development and test sets were created in 15 different languages. XNLI was used to create contradiction detection systems for training in English and predicting in other languages, with good performance. Each example in XNLI is classified as either *contradiction*, *entailment* or *neutral*. However, the NLI resources do not distinguish between different types of contradictions in its annotations. The Spanish language is only available in the development and test sets, and these sets are automatically translated from English. These sets are very small and automatic translation can induce indexable errors in the texts, which may affect the performance of models created from these partitions. In addition, this dataset annotated the semantic relation between two sentences of similar size, which may also affect the performance of models created when both texts are of dissimilar size. This is especially relevant to detecting misleading headlines, which involves semantic matching of headline and body text, and the word count of both texts varies significantly.

C. DATASET 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 [15], *"annotation is also one of the most time-consuming and financially costly components of many NLP research efforts"*. Nowadays, the number of labeled datasets to train is low and data collection is one of the challenges in disinformation research due to the scarce availability of such datasets [39], and this phenomenon is even more pronounced in languages other than English. 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"* [40].

According to the literature consulted, corpus construction in NLP can be approached via several methodologies. Depending on the complexity of the annotation task, the annotation is done completely manually [41], or even completely automatically if the task allows it [42]. However, most of the corpora released for the disinformation task follow an automatic approach for data collection that 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 [43] and [44].

Another type of methodology is crowdsourcing, in which both compilation and annotation can be automatic or manual, such as those introduced by [45], [46], and [47]. This practice enables the bulk outsourcing of multiple labeling tasks, typically with low overall cost and fast completion [48]. It enables 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.

Another possible way to obtain datasets automatically, when dealing with languages other than English, is to perform the automatic translation into the target language from the dataset created originally in English [33], [37], [38]. According to recent research on Chinese language —arguably, the other high-resource language along with English— although translation-based methods and multilingual approaches have an acceptable level of performance, a large margin for improvement exists [49]. One of the main problems involved in performing automatic translation in tasks where semantics play a key role, as in the case of our research, is that semantics may not transfer accurately to the target language and many idiomatic expressions that do not have a direct translation in the source language may not be captured in the target language or vice versa. That is why the generation of a resource from scratch in the original language will always be more accurate and comprehensive, and better able to detect contradictions than resources created from automatic translation, which is likely to propagate the errors of the translation task.

Furthermore, this is especially relevant to the journalistic domain because the language used in a headline is very likely to be catchy and idiomatic so capturing the semantics is more difficult if the classification model has not learned from a natively-created dataset [50], [51]. Thus, it is vital to have native evaluation data for a specific language to measure progress in tasks for this language [52].

For this reason, an important part of our proposal is to obtain comparable natively-created data for the task in the specific language. Apart from this, despite the existence of resources similar to our task — presented in the previous subsection, such as SNLI or XNLI—, the translation of these resources is not the most appropriate option. This is because these datasets annotate semantic relations between similar-sized pieces of text, whereas in our proposal, our goal is to detect contradiction between the headline and the body text.

Furthermore, these resources are only annotated with *contradiction, entailment* or *neutral*, whereas our aim is to define a fine-grained contradiction annotation, distinguishing between different types of contradictions with different semantics that are relevant and useful. The other novelty compared to other semi-automatic methodologies presented in the literature is applying different NLP techniques to automatically create contradictory examples without human intervention, resulting in a more efficient annotation procedure.

III. ES_HEADLINE_CONTRADICTION DATASET BUILDING METHODOLOGY

Complex language models applied in classification tasks need to be trained with quality datasets and with numerous examples. In this sense, we propose a novel semi-automatic dataset building methodology, divided in two annotation phases. Applying the methodology, an extension of the dataset developed in [16] is performed in a more efficient way. The final objective of this dataset creation is using it in automatic contradictions identification that supports the misleading headline detection task for Spanish.

This section will explain the methodology followed to build the dataset. Our research aims to annotate a large-scale contradiction corpus using a semi-automatic approach. First, contradiction theoretical foundations are presented, as well as the different contradictions considered. Then, the dataset building procedure is presented, consisting in a first phase where a manual annotation is done. In the second phase, we automate the annotation of some of the contradiction types so as to obtain a dataset with numerous examples.

A. THEORETICAL FOUNDATIONS OF CONTRADICTION

The task of automatic detection of contradictory information is tackled as a classification problem [53], when two pieces of text are talking about the same fact, within the same temporal frame. Based on the different contradiction definitions in the literature, we establish a dissonance between two pieces of text in three general categories. We define a statement as $s = (i, f, t)$, where i refers to the information provided about fact f occurring at the time t, we classify two pairs of text as:

- *Compatible*: two pieces of text, s_1 and s_2, are considered compatible if, given $s_1 = (i_1, f_1, t_1)$ and $s_2 = (i_2, f_2, t_2)$, the following statement holds true:

$$(i_1 \cong i_2) \wedge (f_1 \cong f_2) \wedge (t_1 \cong t_2) \qquad (1)$$

- *Contradictory*: two pieces of text, s_1 and s_2, are considered contradictory if, given $s_1 = (i_1, f_1, t_1)$ and

$s_2 = (i_2, f_2, t_2)$, the following statement holds true:

$$(i_1 \ncong i_2) \wedge (f_1 \cong f_2) \wedge (t_1 \cong t_2) \tag{2}$$

- *Unrelated*: two pieces of text, s_1 and s_2, are considered unrelated if, given $s_1 = (i_1, f_1, t_1)$ and $s_2 = (i_2, f_2, t_2)$, the following statement holds true:

$$f_1 \neq f_2 \tag{3}$$

Thus, a news item is classified as *contradictory* when given the same fact[7]. within the same time frame, the related information is incongruent in the two news items being considered.

In practice, references to the time variable are not usually found in the verification of semantic relations between news items. To this end, an abstraction is made, anticipating that if two texts are being compared for the purpose of searching for contradictions then they belong to the same time frame.

Similar to the FNC-1 dataset, the proposed dataset annotates the semantic relationship between headlines and body text. However, unlike the FNC-1 dataset, where the semantic relationship was defined in terms of the stance between the two pieces of text, in our dataset the relationship is defined in line with the definition of contradictions aforementioned. Following this definition, a headline and body text could be classified as *compatible*, *contradictory*, or *unrelated*.

In the case that the relationship is of type *contradictory*, we follow the classification of the different types of contradictions proposed by [13]. Our research is focused on a subset of them, more specifically: *negation*, *antonym*, *numeric/date*, *factive*, and *structure*. The subsequent definitions of contradiction types were arrived at by analysing two related sentences that exhibit contradiction:

1) *negation*: the main event in one of the sentences analysed is negated, causing the sentence to completely change its meaning. Negation marks (no, none, never, etc) are used.
2) *antonym*: the two main events in each sentence are antonyms, turning two semantically compatible sentences into contradictory ones.
3) *numeric/date*: there are differences between parts of sentences expressing numeric data or dates, making the sentences contradictory.
4) *structure*: the structure of one of the sentences is not compatible with the other. The named entity performing an action is different from the one found in the other sentence, or named entities in a sentence are interchanged.
5) *factive*: one of the sentences uses a factive verb so the writer shows commitment to the truth of the proposition expressed, whereas the other sentence uses non-factive verbs, this is, the writer does not grant factual status to the proposition, not that s/he considers the proposition to be false. Factive verbs are verbs that take a clause as a complement and introduce a presupposition that the complement clause is true [54].

In our research, we consider that the annotation of each type of contradiction could be beneficial to generating the explanation of decisions made by future systems developed on this dataset.

B. DATASET BUILDING

In order to explain the process of constructing the dataset, the following main phases were defined: planning, first annotation phase, and second annotation phase. Each of the proposed phases is explained below.

1) PLANNING

The first step is to choose a reliable data source to extract the news. In this case, the news agency EFE[8] was used because it is known for its neutrality in its publications [55]. Being a news agency, they act as information providers, between events and the media. Moreover, EFE is the main news agency that feeds most of the Spanish media, and is therefore an appropriate news source without the possible bias of the different digital media publishers [56].

Secondly, news items, consisting of the headline, the body text and the date of the news item, are extracted. In this research, the extracted news items belong to the political and economic domains. These two domains provide a lot of numerical and factive events, as well as very well known named entities, such as organizations or political parties. This allows for easier manipulation of headlines than in the case of news that is broader in scope, such as social or cultural news. From the extracted news item, since they come from the news agency, it is assumed that the headline and the body text of the news items are *compatible*, although their relationship is subsequently verified in the annotation process.

Finally, a web crawler is implemented using the Python library BeautifulSoup[9], which downloaded a total of 25,945 news items between January 2019 and March 2021.

In addition, an annotation guideline was developed, explaining in detail the procedure to be followed to modify headlines and create ones that fit the type of contradiction in line with the definition of contradiction presented by [13]. This guideline is available at the following link from Zenodo[10].

2) FIRST ANNOTATION PHASE

The first annotation phase aimed to develop a preliminary version of the dataset. The first version did not have such a high cost since a fraction of the downloaded news was annotated.

In the first annotation phase, a randomly selected subset of the downloaded news (7,403) was chosen to develop the first

[7] The same fact in two different news items could be expressed with different event mentions

[8] https://www.efe.com/ (accessed 12 July 2023)

[9] Documentation available at https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed 1 February 2023)

[10] https://zenodo.org/badge/latestdoi/344923645 (accessed 1 February 2023)

version of the dataset. These news items represent 28.53% of the total news items (25,945). The 7,403 news items were divided randomly into three sets. In the first set, 2,508 (33.87%) headlines were manually modified to include all types of contradiction. In the second set, 2,396 (32.36%) news items were added to the first part to be manually annotated as *compatible* or *contradictory*, but without being modified. Finally, the third set, 2,499 (33.75%) was used for the generation of examples of type *unrelated*. This split was performed pursuing the objective of obtaining a dataset with balanced classes. The scikit-learning library is used to partition news randomly.

This phase was divided into three main tasks: **manual modification of news headline**; **classification of headline and body text relationship**; and, **random pairing of headlines to their non-original body text**. Each task is explained below:

1) **Manual modification of news headline**: the aim of this task is to modify the headline so that it contradicts the body text, by including simple modifications to its semantics. The changes to the headline along with some examples are shown below:
   - *negation (neg)*: This alteration consists of negating the news headline by including a negation indicator in a specific position in the sentence.
     a) Original headline: "El comité de empresa **debatirá** mañana la propuesta final de Alcoa" *(Union representatives **will discuss** Alcoa's final proposal tomorrow).*
     b) Contradictory headline: "El comité de empresa **no debatirá** mañana la propuesta final de Alcoa" *(Worker's council **will not debate** Alcoa's final proposal tomorrow).*
   - *antonym (ant)*: This transformation consists of replacing the verb of the main event in the headline with an antonym.
     a) Original headline: "El Gobierno se compromete a **subir** los salarios a los empleados públicos tras los comicios" *(The Government pledges to **raise** public employees' salaries after the elections).*
     b) Contradictory headline: "El Gobierno se compromete a **bajar** los salarios a los empleados públicos tras los comicios" *(Government pledges to **cut** public employees' salaries after the elections).*
   - *numeric/date (num)*: This modification consists of changing the numbers, dates that appear in the headline.
     a) Original headline: "La economía británica ha crecido un **3%** menos por el brexit, seg ún S&P" *(UK economy has grown by **3%** less due to Brexit, says S&P).*
     b) Contradictory headline: "La economía británica ha crecido un **5%** menos por el brexit, seg ún S&P" *(UK economy has grown by **5%** less due to Brexit, says S&P).*
   - *structure (str)*: This modification consists of changing the position of one named entity to another or substituting named entities in the sentence.
     a) Original headline: "**Arvind Krishna** sustituirá a **Ginni Rometty** como consejero delegado de IBM" *(**Arvind Krishna** will replace **Ginni Rometty** as IBM's CEO).*
     b) Contradictory headline: "**Ginni Rometty** sustituirá a **Arvind Krishna** como consejero delegado de IBM" *(**Ginni Rometty** will replace **Arvind Krishna** as IBM's CEO).*
   - *factive (fac)*: This transformation consists of replacing the main event verb with a non-factive verb construction or vice versa.
     a) Original headline: "Isuzu y Volvo **pactan crear** una alianza estratégica en camiones pesados" (Isuzu and Volvo **agree to create** a strategic alliance in heavy duty trucks).
     b) Modified headline: "Isuzu y Volvo **crean** una alianza estratégica en camiones pesados" (Isuzu and Volvo **create** a strategic alliance in heavy duty trucks)

   These alterations change the semantic meaning of the headline, making it *contradictory* to the original one and the body text. The annotation process was carried out by two independent annotators who were trained by an expert annotator. The first set of news items (2,508) was used. This first phase generated *negation (neg), antonym (ant), numeric/date (num), and structure (str)* contradictions.

2) **Classification of headline and body text relationship**: The semantic relationship between the headline and the body text was annotated in two steps. The first step consisted of annotating the information as *compatible* or *contradictory*. In the second step, when the headline-body text relationship was annotated as *contradictory*, the type of contradiction (*negation*, *antonym*, *numeric/date* or *structure*) was also annotated. This task involved four annotators trained to detect semantic relations between pairs of texts.

3) **Random pairing of headlines to their non-original body text**: The third set of news items (2,499) reserved at the beginning of this phase was used to generate *unrelated* examples. The headline was separated from the corresponding body text and all headlines were randomly assigned to their non-original body text. This task is similar to the one used for obtaining unrelated examples in the FNC-1 dataset [57]. This step was done automatically without the intervention of annotators.

As can be seen from the explanation of the tasks in this phase, the first two tasks required the intervention of human annotators. In contrast, the third task was performed automatically. In this first annotation phase, the contradiction type *factive* was not annotated because it is a contradiction that requires more effort to modify headlines.

3) SECOND ANNOTATION PHASE

The second annotation phase aims to increase the number of examples by annotating them automatically when possible. Furthermore, the number of examples to be annotated per contradiction type is planned with the main aim of maximising examples of more complex contradictions such as *structure* and *factive* contradictions. The annotation guideline followed in this second phase is the same as in the first phase.

News items that were not used in the first annotation phase (18,542) were used in this phase. The second phase does not include the task **Classification of headline and body text relationship** because headlines without modifications were *compatible* with the body text, as demonstrated in the validation process of the first phase. The second phase consists of three tasks: **Automatic modification of headline**; **Manual modification of news headline**; and, **Random pairing of headlines to their non-original body text**.

1) **Automatic modification of headline**: In this task, an automatic headline modification mechanism creates contradiction types without the intervention of human annotators. Based on the experiences gained from the manual annotation carried out in the previous phase, and following the same annotation guidelines[11], a pipeline is created that produces three types of contradictions (*antonym*, *numeric/date* or *structure*). The pipeline[12] was created by making use of the Spacy library, which specialises in NLP. Each headline is pre-processed by the pipeline. Figure 1 shows the internal components of the pipeline[13].

   Spacy makes it possible to create pipelines in a simple way, reusing available components and developing and integrating your own ones. Each of the components used and created are explained below:

   a) *tokenizer*: segments the text into tokens.
   b) *tagger*: assigns the speech tags.
   c) *parser*: performs an analysis of dependencies between tags.
   d) *ner*: detects the named entities.
   e) *num*: detects the ways of expressing a number (ordinal, cardinal, and digits).
   f) *date*: detects dates included in a text (days, months, years, etc.).
   g) *mod*: modifies the headlines using the annotation of the previous components. This component is responsible for selecting the type of modification based on the specified priority, and the total number of annotated examples of each type of contradiction. The order of priority is *structure*, *numeric/date*, and *antonym*.

The last three components of the pipeline were created for this research to enable the automatic modification of the headlines.

The *num* and *date* components of the pipeline have been created using a rule-based procedure defined by the Spacy Matcher component. These rules have made it possible to detect complex patterns (like first semester, third quarter, etc).

In the last component of the modification pipeline (*mod*), depending on the headline to be modified, one type of contradiction or another can be included. For example, in the case of an *antonym* contradiction, if an antonym is not found for the main verb, this headline is not modified with this type of contradiction, and the pipeline will try to include another type.

In order to modify the headlines automatically with each type of contradiction, a series of decisions were made to guarantee the integrity of the examples created as well as their diversity. In the case of the *antonym* contradiction, we used a publicly available resource[14] that finds antonyms of words. Next, the morphological annotation of the verb is used to conjugate the antonym in the same mode, tense, number, and subject-verb agreement. A public resource[15] was used to perform this conjugation. In the case of the *numeric/date* and *structural* modification task, the modification was done according to a series of parameters. Once the type of element to be modified was detected, a set of defined rules was applied to restrict the modification to another similar type of element within a valid range value. In the case of numeric/date-type modifications, for example, if the original headline to be modified includes the numerical value (15%), the automatic modification pipeline would never transform this value to a non-possible one (i.e. 400.00%). In the case of structure-type modifications, for example, if the element to be changed is a person, this element of the headline would never be modified to a location.

In Figure 2 there is an example of the structure modification[16] (the original example in the dataset in Spanish is: Input headline: *Corea del Sur retira a Japón de su lista de socios comerciales preferentes*; Modified headline: *Japón retira a Corea del Sur de su lista de socios comerciales preferentes*). This figure only shows the elements of the sentence implied in the modification process for clarity purposes. First, the well-known components of the Spacy pipeline (*tokenizer*, *tagger*, *parser*) are executed. Next, the *ner* component detects the named entities and their types, followed by the

[11] https://zenodo.org/badge/latestdoi/344923645 (accessed 1 February 2023)

[12] Implementation available at https://github.com/rsepulveda911112/contradiction_spacy_pipeline

[13] Based on the figure taken from https://spacy.io/usage/processing-pipelines/ (accessed 1 February 2023).

[14] https://www.wordreference.com/sinonimos

[15] https://conjugador.reverso.net/conjugacion-espanol.html

[16] Figure example is presented in English for better understanding.

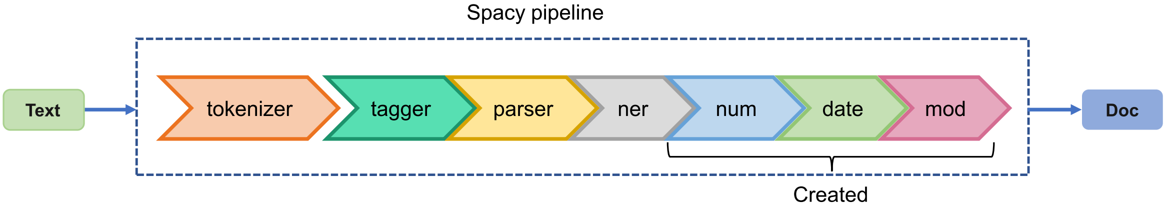

FIGURE 1. NLP pipeline created using the spacy library.

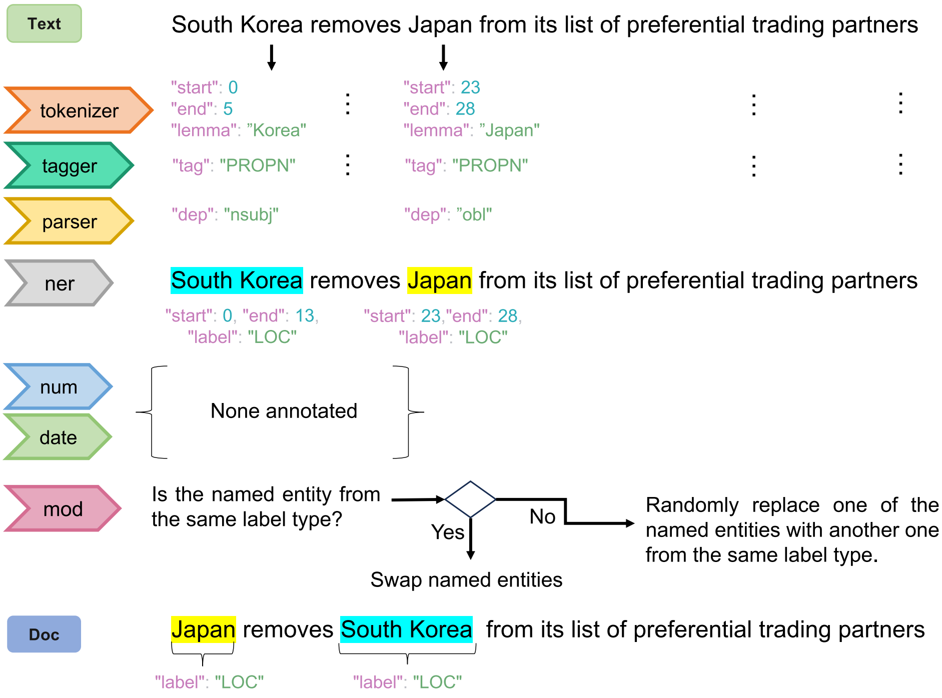
Text
South Korea removes Japan from its list of preferential trading partners
tokenizer
"start": 0
"end": 5
"lemma": "Korea"
"start": 23
"end": 28
"lemma": "Japan"
tagger
"tag": "PROPN"
"tag": "PROPN"
parser
"dep": "nsubj"
"dep": "obl"
ner
South Korea removes Japan from its list of preferential trading partners
"start": 0, "end": 13,
"label": "LOC"
"start": 23,"end": 28,
"label": "LOC"
num
date
None annotated
mod
Is the named entity from the same label type?
Yes
No
Randomly replace one of the named entities with another one from the same label type.
Swap named entities
Doc
Japan removes South Korea from its list of preferential trading partners
"label": "LOC"
"label": "LOC"

FIGURE 2. Example of structure modification.

mod component that verifies named entities types and swaps them or randomly replaces one of the named entities for another of the same label type. In the case of this example, due to the fact that it is a structure contradiction, the components *num* and *date* do not return any annotation.

In addition, the values introduced for the modified elements are random, which enriches the modification approach. Of the total number of news items, a batch of 6,778 was automatically annotated with *antonyms* (1,952), *numeric/date* (2,150) and *structure* (2,676).

2) **Manual modification of news headlines**: This task has the same objective as the same task in the previous phase. However, this task annotates the contradiction type *factive* as it is highly complex to include in the automatic modification pipeline. Additionally, some headlines were modified by introducing the *negation* contradiction, which in experimental tests generated inconsistent examples in the automatic version of the headline modification. Of the total number of news items, a batch of 808 was manually annotated between *negations* (241) and *factive* (567).

3) **Random pairing of headlines to their non-original body text**: This task is exactly the same as the one described in the first annotation phase. In total, 3,610 headline-body text pairs are used to generate the examples *unrelated*.

IV. ES_HEADLINE_CONTRADICTION: SPANISH MISLEADING HEADLINES DATASET

Due to the fact that the dataset was developed in two phases of annotation, there are two versions of the dataset. In the first phase, four types of contradictions are annotated (*negation*, *antonym*, *numeric/date* and *structure*) and in the second phase, the contradiction *factive* is added.

A. FIRST VERSION OF THE DATASET

Following the first annotation phase, the first version of the developed dataset consisted of 7,403 news items, of which 2,431 were annotated as *compatible*, 2,473 as *contradictory*,

TABLE 1. Distribution of classes in each partition of the first version of the dataset.

Set	Compatible	Contradictory	Unrelated
Training	1,703	1,733	1,755
Test	728	740	744
Total	2,431	2,473	2,499

TABLE 2. Distribution by type of contradictions in the first version of the dataset.

Set	Neg	Ant	Num	Str
Training	674	552	430	77
Test	287	236	184	33
Total	961	788	614	110

and 2,499 as *unrelated*. This represents a balanced dataset with three main classification elements. The dataset was divided into training and test partitions. The distribution for each partition is shown in table 1.

There are far fewer examples of the class *str* compared to the rest, suggesting that it is the most complex one to generate in headlines. In this phase, the annotation is manual and from the assigned set of headlines. The annotator analysed each headline and decided from among the types of contradictions the most appropriate one to apply. In the manual annotation step, the decision of the human annotator in so far as selecting the most appropriate contradiction as being closest to a real possible scenario is considered valuable.

The training and test partitions were created by allocating 70% of the dataset to training and 30% to test. As can be seen in table 2, the dataset contains examples of each type of contradiction.

1) VALIDATION OF FIRST VERSION OF THE ANNOTATION

Due to the particularities of the annotation process for this dataset, it was necessary to validate the tasks **(1) Manual modification of news headline** and **(2) Classification of headline and body text relationship**. For the first task, a validation was performed by an expert involved in the creation of the annotation guideline.

For the second task, two different validations were performed. Firstly, an inter-annotator agreement was made. 200 examples (4% of the pairs annotated as *compatible* and *contradictory*) were randomly extracted to perform validations on the first version of the dataset. For the sample *contradictory*, we validated a balance between each type of contradiction in the extracted examples. Finally, an analysis of the different scenarios, in which errors occur in the classification of the semantic relation, was presented and discussed.

- **Expert validation**: For the first task, it was not possible to reach an agreement between annotators since this task consisted of modifying the headlines and the possible variants may be unmanageable or tend to infinity. In this case, an expert annotator performed a manual review of the modified headlines in order to detect inconsistencies with the indications in the annotation guideline. We observed that only 2% of the examples analysed present inconsistencies with the annotation guideline, which corroborates the validity of the process developed for this task.
- **Inter-agreement between two annotators**: In order to measure the quality of the second annotation task, an inter-agreement between two annotators was made. They independently annotated 200 examples between *compatible* and *contradictory*, calculating an annotation agreement index. Cohen's kappa index was used to calculate annotation agreement (a common index in annotation validation processes between two annotators) [58]. A Cohen's kappa of 0.83 was obtained, representing a high value of agreement between two annotators, validating the annotation process. In cases where there was no agreement (or coincidence), a consensus process was developed between the annotators, which highlighted that there were erroneous interpretations of the annotation guideline. We observed that most of the problems in annotation were related to annotating the specific type of contradictions.
- **Validation of annotation errors**: After annotating the semantic relationship between headlines and body text (task 2), we found that 2,524 news items were annotated as *contradictory*, compared to the 2,508 headlines that were modified in task 1. As shown in table 3, there is a marked difference between the manually modified headlines (row 1) and those annotated by type of contradiction (row 2).
  In this last validation, three situations were analysed:
  - In the first analysis of this validation process, we found that the annotators annotated most of the original headlines as *compatible* (only 58 news items were annotated as *contradictory*) which underscores the reliability of the EFE agency. Two annotators reviewed the 58 news items and found that the annotators made incorrect interpretations of the annotation guideline or classified the relation between the headline and body text incorrectly. After this process, we re-annotated them as *compatible*. It is important to clarify that most of these mistakes were found in headlines with more than one named entity or those including figures, which led to the misclassification of 28 *numeric/date*, 20 *structural*, 7 *antonym*, and 3 *negation* contradictions.
  - The second analysis was to verify the headlines that were modified but were annotated as *compatible*. 36 news items were misclassified and should have been classified as follows: 24 *negation*, 5 *antonym*, 2 *numeric/date*, and 5 *structure*. After analyzing each case in detail, it was observed that 31 headlines were not properly modified in task 1. Most of them presented concordance problems or the

TABLE 3. Headlines modified (task 1), headlines annotated (task 2), and final annotation (after validation).

	Neg	**Ant**	**Num**	**Str**	Total
Headlines modified (in task 1)	985	793	616	114	2,508
Headlines annotated as contradictory (in task 2)	949	809	644	122	2,524
Final annotation (after validation)	961	788	614	110	2,473

modifications did not produce any type of contradiction. For these cases, the original headline was restored and they were annotated as *compatible*. Furthermore, 5 headlines that had been modified correctly in task 1 were subsequently incorrectly annotated in task 2 as *compatible*. Thus, these news items were re-annotated as *contradictory* and given their appropriate contradiction type (3 *structure* and 2 *negation*).

- Finally, 86 news items were incorrectly classified for type of contradiction, as shown in table 4. Most of the errors were found among the *num* and *str* contradiction types. The next most common errors were for items classified as *ant* when they should have been classified as the *neg* type. This contradiction classification error is less important because the news items were correctly classified as *contradictory*. These errors were rectified and the correct type of contradiction was annotated. This analysis also exposes the possible overlap between different types of contradictions.

TABLE 4. Confusion matrix.

	Neg	**Ant**	**Num**	**Str**
Neg	—	**13**	**7**	6
Ant	3	—	5	2
Num	**7**	6	—	**12**
Str	5	5	**15**	—

The validation statistics can be consulted in the Zenodo repository of annotation guidelines shown above.

These validations have demonstrated that the majority of news items from the EFE agency were classified as *compatible* (97.6%). This insight provides the basis for the enrichment of the annotation methodology proposed next. In this sense, some automatic modification of headlines can be introduced, thereby making it unnecessary to annotate the semantic relation between them as *compatible* and *contradictory*. The original headlines will be classified automatically as *compatible* and the modified ones as *contradictory*, thus saving time in the annotation process. If many contradictions were to be found in the original headlines, the semantic relationship would be annotated.

B. SECOND VERSION OF THE DATASET

After performing the second annotation phase, the second version of the developed corpus was obtained, consisting of 18,542 news items, of which 7,346 have been annotated as *compatible*, 7,586 as *contradictory* and 3,610 as *unrelated*. In this version, three contradictions were annotated automatically (*str*, *ant* and *num*) and two manually (*neg* and *fac*). As in the previous version, the dataset was divided into training and test partitions. The distribution of classes is shown in table 5.

TABLE 5. Distribution of classes in each partition of the second version of the dataset.

Set	**Compatible**	**Contradictory**	**Unrelated**
Training	5,142	5,308	2,527
Test	2,204	2,278	1,083
Total	7,346	7,586	3,610

Finally, table 6 shows the distribution of classes by contradiction type. This version shows a less uniform class distribution because it was planned to maximise some contradiction types with fewer examples.

TABLE 6. Distribution by type of contradiction in the second version of the dataset.

Set	**Neg**	**Ant**	**Num**	**Str**	**Fac**
Training	168	1,366	1,505	1,873	396
Test	73	586	645	803	171
Total	241	1,952	2,150	2,676	567

As can be seen in the table and if we compare it with the data from the first phase (table 2), all the types with automatically generated examples (*ant, num, str*) have increased significantly. The increase in one of the most complex types (*str*) is particularly remarkable, being able to go from 110 manual examples to 2,676 automatic ones.

1) VALIDATION OF THE SECOND VERSION OF THE ANNOTATION

To evaluate the tasks performed in the second version of the annotation, two validations were carried out. The first validation corresponds to the headline modification tasks —**(1) Automatic modification of headline** and **(2) Manual modification of news headline**— that have been verified by an expert. The second validation consisted of an inter-annotator agreement to ensure the validity of the dataset.

We randomly selected 590 examples (representing nearly 8% of modified headlines) for the first validation, and 200 examples (between *compatible* and *contradictory* pairs) for the second validation. In the *contradictory* pairs selected for both validations, there is a balance between headlines modified automatically and manually.

- **Expert verification**: Manual verification by an expert is performed. The manually modified examples are in line with the annotation guideline. In a first validation, the automatically modified examples produced a high number of concordance problems in the headlines, around 15%, which made it necessary to modify the spacy pipeline again. A second validation significantly reduced these errors to 4%. Most of them were found in the contradictions of *ant* and *num*. This validation corroborates the validity of the manual and automatic headline modification process.
- **Agreement between annotators**: The second annotation phase does not use an extra task to annotate the semantic relation. It builds on the experience gained in the previous phase, where very few original news headline-body text pairs were annotated as *contradictory*, evidencing the quality of the news from the chosen source. However, once the headline modification tasks have been validated, an additional inter-annotation agreement between two annotators manually classifying the headline-body text relationship was performed. Cohen's kappa index is also used with a result of 0.79, which means substantial agreement for the dataset annotation.

C. ANNOTATION PROCEDURE DETAILS

For the manual phase, a total of six annotators were used. The annotators were linguistic experts specialized in NLP and all of them native Spanish speakers. No special prior proficiency or expertise is required in the annotation process as this avoids possible bias in it. Only the ability to establish the appropriate contradiction following the annotation guide is necessary. Two of the annotators were in charge of the headline modification process and the other four performed the annotation of the semantic relationship between the headline and the body text. Each type of annotator was specifically trained to carry out their process using the annotation guideline. The annotators in charge of annotating the semantic relationship between the headline and the body text did not know the process involved in modifying the headlines, thereby mitigating the risk of bias in the methodology.

For the semi-automatic phase, only two annotators were required, one of them was in charge of the headline modification and the other was responsible for annotating the semantic relation. This was done, as in the manual phase, to maintain total independence between the annotation tasks. Annotators did not use any specific user interface but a general notepad application.

Table 7 indicates the number of news items annotated both manually and semi-automatically and logs the time taken for each phase. The data in the table enable us to assess the benefits of applying the methodology for creating the dataset, i.e., whether efficiency in terms of dataset construction is improved without compromising the quality of the dataset.

As observed in the table 7, the semi-automatic phase is approximately 150.48% larger than the first phase in terms of the amount of news generated, whereas the time spent on news generation in this second phase has been reduced by approximately 88.8% compared to the manual process, thereby allowing us to improve the models that learn with this larger input. In addition to reducing the time spent on generating a substantially larger news set, the number of annotation resources is also reduced because the task has been performed by fewer annotators (only 2 in this case) without hardly compromising the quality of the dataset, as can be seen in the value of the inter-annotator agreement.

D. CONSOLIDATION OF THE DATASET: ES_HEADLINE_CONTRADICTION DATASET

After performing the two annotation phases, and merging the dataset obtained in the first phase (Dataset_V1) with the dataset obtained in the second phase (Dataset_V2), the ES_Headline_Contradiction dataset was obtained. As explained before, in the first phase, four types of contradictions were annotated (*negation*, *antonym*, *numeric/date* and *structure*) whereas in the second phase, the contradiction *factive* was added.

Completing the two annotation phases delivered enough examples to have a relevant dataset for the task of misleading headline detection. In this sense, table 8 shows the distribution of classes of the consolidated version of the dataset.

Additionally, table 9 indicates the distribution by contradiction type. It is clear that the contradictions *ant*, *num* and *str* contain a similar number of annotated examples. However, the contradiction type *fac* has very few annotated examples, which may affect the performance of a future contradiction detection model for this contradiction type. Both the final version and partial versions are available at this link to Zenodo[17].

In order to illustrate the percentage distribution by type of contradictions the following graph is presented (figure 3).

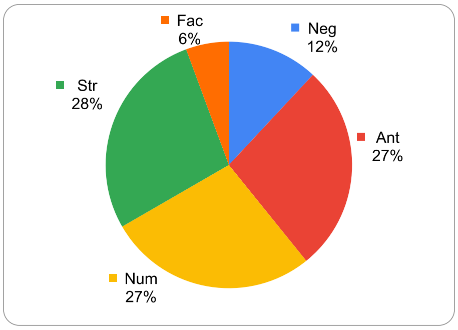

FIGURE 3. Percentage distribution of contradiction types in ES_Headline_Contradiction dataset.

Finally, in table 10, the statistics related to the ES_Headline_Contradiction dataset are displayed. The average word count of *tokens* in the headline and the body text of

[17] https://zenodo.org/badge/latestdoi/344923645 (accessed 1 Feb 2023)

TABLE 7. Efficiency of the annotation for each phase of the dataset creation.

Phase	No. news	Total Annotation time (hours)	No. Annotators	Cohen's kappa
First (manually)	7,403	276	6	0.83
Second (semi-automatic)	18,542	31	2	0.79

TABLE 8. Distribution of classes in each partition of the dataset ES_Headline_Contradiction.

Set	Compatible	Contradictory	Unrelated
Training	6,845	7,041	4,282
Test	2,932	3,018	1,827
Total	9,777	10,059	6,109

TABLE 9. Distribution by type of contradictions in the dataset ES_Headline_Contradiction.

Set	Neg	Ant	Num	Str	Fac
Training	842	1,918	1,935	1,950	396
Test	360	822	829	836	171
Total	1,202	2,740	2,764	2,786	567

TABLE 10. Dataset Statistics Overview.

Set	Headline word count (Average)	Body text word count (Average)
Training	13.6	545.2
Test	13.6	553.8
Total	13.6	547.8

the news item is calculated. For this purpose, the tokenizer of the Spacy library is used, which operates at word level.

V. EXPERIMENTS

A set of experiments was conducted to demonstrate that a classification system is able to detect contradictions that indicate misleading headlines from the proposed dataset. Figure 4 shows the internal structure of the classification system used.

The classification system receives as input the headline and the body text. These inputs are concatenated to be subsequently processed by a language model that is responsible for encoding the semantic relationship between them. The output of this language model is passed to a classification neural network (feed-forward fully-connected type), which is in charge of classifying the type of relationship.

This classification system allows fine-tuning on a specific task. For this purpose, the system is put into fine-tuning mode and the training examples of our dataset are used to adjust the weights of the language model as well as those of the neural network classifier. To perform the fine-tuning process, hyperparameter settings defined at the end of this section are used.

To perform the experiments we selected language models that implement the transformer architecture, which obtain state-of-the-art results in the main tasks within NLP. For this, we used two general language models in Spanish as well as other specific language models. These models allow us to evaluate the benefits of transfer learning when using these models in our task. More specifically, considering the semantic relationship between headline-body text in *compatible*, *contradictory*, and *unrelated* as annotated in our dataset, a strong relation with the NLI task is found, where the semantic relationship between two sentences is annotated as *entailment*, *contradiction*, and *neutral*. There are differences between our dataset and those of the NLI task (SNLI, MultiNLI, and XNLI) such as the word count of the annotated texts and the annotated semantic relations. However, the NLI task datasets have been used to train specific and multilingual language models, so it was decided to test some trained NLI language models to address this task.

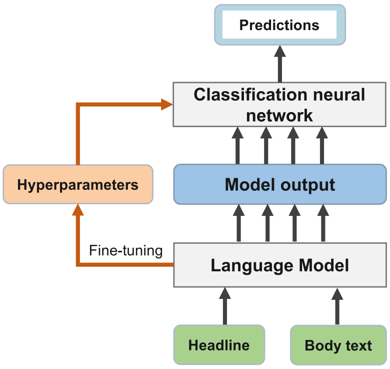

FIGURE 4. Internal structure of classification system.

Therefore, finally, to perform our experiments, we selected four pre-trained models: two general language models in Spanish (**BETO** and **RoBERTa-base-bne**), and two models trained for the NLI task —a specific language model in Spanish (**Spanish_NLI model**), and a multilingual model (**Multilingual_NLI model**).

All these models use the Transformer architecture proposed by [59]. A detailed explanation of each model is presented below:

- **BETO model**[18] was obtained based on the BERT model [60] but with a set of optimizations similar to the RoBERTa model [61]. BETO was trained with texts in Spanish from Wikipedia and the OPUS[19] project. BETO

[18] Available to download in https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased (accessed 16 Jan 2023).

[19] Dataset in parallel with more than 90 languages, the Spanish-English pair being the most representative with 36 million sentences [62], https://opus.nlpl.eu/ (Accessed 16 Jan 2023).

has 12 self-attention layers with 16 attention heads each, using 768 as the hidden size [63].

- **RoBERTa-base-bne model**[20] used the same architecture and optimization as the RoBERTa model [61] with 12 layers, 768-hidden, 12-heads, and 125M parameters. In this case, this model was trained with a total of 570GB of clean Spanish text to obtain a Spanish pre-trained model [64].
- **Spanish_NLI model**[21] based on BERT model was only trained with the Spanish partition of the XNLI dataset. This model has 12 self-attention layers with 12 attention heads each, using 768 as the hidden size.
- **Multilingual_NLI model**[22] was built using as a base the DeBERTa model [65] and was fine-tuned for the NLI task using machine translation to 27 languages (including Spanish). To perform fine-tuning, the following datasets were used: MultiNLI, ANLI [66], WANLI [67], and LingNLI [68]. The XNLI dataset was used to evaluate the model obtained. This model has 12 self-attention layers with 12 attention heads each, using 768 as the hidden size [69].

Some experiments conducted consisted of fine-tuning the four models for classification tasks by making use of the training set of the ES_Headline_Contradiction dataset. The experiments were carried out using the Simple Transformers[23] library. The hyperparameter settings for all experiments are: maximum sequence length of 512, batch size of 4, training rate of 2e-5, and training performed over 3 iterations.

A. EXPERIMENTS TYPE DESCRIPTION

The experiments use the training set of the ES_Headline _Contradiction in the event that fine-tuning is done, and they are evaluated by predicting the test set. The experiments could be replicated using this Github repository[24]. The following experiments were proposed:

1) **Prediction of all classes**: aims to predict the main classes of the dataset *compatible*, *contradictory*, and *unrelated*.
2) **Detection of specific types of contradictions**: uses only the examples of type *contradictory* from the dataset described in table 9 to detect each type of contradiction.
3) **Detection of contradictory vs compatible headlines**: aims to evaluate if off-the-shelf NLI models could serve as baselines for the task. In this case, the **Spanish_NLI model** and **Multilingual_NLI model** are used to predict only the classes *contradictory* and *compatible* since the *unrelated* class is not classified in these NLI models.
4) **Comparison between dataset versions**: aims to evaluate the performance of the best language model using the training and test set of each dataset version. The two preliminary versions of the dataset and then the whole version are used to carry out this experiment.

B. EVALUATION METRICS

To evaluate the performance of the proposal, the F_1 class-wise and macro-averaged F_1 (F_1m) metrics [57] are also used to address the imbalance among the less represented classes. F_1 can be formulated as follows:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precison} + \text{Recall}} \tag{4}$$

F_1 combines precision and recall in an harmonic mean. Precision measures the accuracy of positive predictions, while recall measures the completeness of positive predictions. F_1m is computed as the mean of those per-class F scores. The advantage of using these metrics is that they are not affected by the size of the majority class.

VI. RESULTS AND DISCUSSION

This section presents the results obtained in each of the experiments described in Section V. Values are expressed as percentages.

A. PREDICTION OF ALL CLASSES

This experiment aims to predict the three (3) classes defined above —*contradictory*, *compatible*, and *unrelated*. Thus, it was performed on the entire dataset using the three language models chosen. Table 11 presents the results.

After performing the fine-tuning on the ES_Headline_Contradiction, the best results were obtained by predicting the test set with the system Multilingual_NLI_model (*fine-tuned*). This system was obtained from the model that was previously trained for the NLI task (Multilingual_NLI model). However, the specific language model (Spanish_NLI model), also trained for the NLI task, obtained the worst results on the F_1m and the F_1 per class metrics. The results generated by the system that uses the Spanish_NLI model are to be expected because it was only trained on the development set of the XNLI dataset. This set is considered too small for a complex task such as NLI compared to the datasets used to train the Multilingual_NLI model. In addition, the developed set of XNLI in Spanish was created using automatic machine translations which can produce unexpected errors in the texts, affecting the performance of systems created with this model. For this reason, we consider that the Spanish_NLI model has some limits when used for fine-tuning in specific tasks.

For its part, the BETO (*fine-tuned*) and RoBERTa-base-bne (*fine-tuned*) systems, trained from the general language model in Spanish (BETO and RoBERTa-base-bne respectively) obtained remarkable results, only surpassed in 1.35%

[20] Available to download in https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne (accessed 29 May 2023)

[21] Available to download in https://huggingface.co/Recognai/bert-base-spanish-wwm-cased-xnli (accessed 16 Jan 2023)

[22] Available to download in https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (accessed 16 Jan 2023)

[23] Documentation available at https://simpletransformers.ai/ (accessed 16 Jan 2023)

[24] https://github.com/rsepulveda911112/ES-Contradiction-baseline

TABLE 11. **Results obtained in experiment 1: Prediction of *compatible*, *contradictory*, and *unrelated*.**

		F_1 Score (%)		$F_1m(\%)$
Systems	**Compatible**	**Contradictory**	**Unrelated**	
BETO (*fine-tuned*)	91.60	91.66	99.58	94.28
RoBERTa-base-bne (*fine-tuned*)	92.20	92.33	**99.69**	94.74
Spanish_NLI_model (*fine-tuned*)	86.20	85.59	99.28	90.36
Multilingual_NLI_model (*fine-tuned*)	**93.73**	**93.67**	99.50	**95.63**

and 0.89% of F_1m by the multilingual system. The RoBERTa-base-bne model surpassed the result of the BETO model, confirming the findings shown by [64]. In addition, these results confirm that fine-tuning models on a specific task similar to the one being addressed are more advantageous than starting from a model trained on generic tasks such as the BETO and RoBERTa-base-bne model.

The results obtained in the class *unrelated* to predict the test set indicate that the systems are capable of detecting this class, with high performance, corroborating the results obtained in the literature on this type of semantic relationship between texts [27]. With respect to the other two classes, the systems achieved remarkable results, but there is room for improvement. A possible option to improve these results could be to include external knowledge. A future line of work would consist of including resources that detect antonyms and synonyms in line with [70] so as to improve the results of the *contradictory* class. In addition, including syntactic and semantic information could improve the detection of other more complex contradictions, such as *structural* and *factive*, without the need for such high cardinality datasets.

B. DETECTION OF SPECIFIC TYPES OF CONTRADICTIONS

This experiment aims to analyse the detection capability for each contradiction type. The results are presented in Table 12, which shows the results obtained exclusively for the detection of contradiction types, not including *compatible* or *unrelated* headlines in the experiment.

Similar to the experiment in the previous section, the multilingual system —Multilingual_NLI_model (*fine-tuned*)— that was pre-trained for the NLI task performed better than other systems. The F_1m results achieved by the multilingual system (92.55%) are significant compared with Spanish_NLI_model (*fine-tuned*) that achieved 84.89% and BETO (*fine-tuned*) which attained 89.82%. In contrast, RoBERTa-base-bne was only surpassed by 0.23% percentage points by the Multilingual_NLI_model, showing the potential of this model after fine-tuning.

The worst results were obtained in the *fac* class, which is considered more complicated to detect compared to the other contradictions [13], and there are very few annotated examples compared to the other classes (only 6% of the total contradictory examples as shown in figure 3). Interestingly, the contradiction type *neg* with far fewer examples than the classes *ant*, *num* and *str* (12% of the total contradictory examples) achieved remarkable results (96.49% of F_1), which indicates that deep learning language models are capable of easily learning to detect this type of contradiction. However, as indicated in the automatic modification process, automatically adding a negation marker with a rule-based system generates numerous cases that cause the headline to lose concordance or consistency. For this reason it was decided not to include *negation* contradiction in the automatic modification process.

C. DETECTION OF CONTRADICTION VS COMPATIBLE HEADLINES

Based on the aforementioned similarity between our task and the NLI task, it was decided to use two off-the-shelf NLI models with the aim of evaluating performance in predicting only the classes of our dataset. NLI is a task for determining whether the given "hypothesis" (H) and "premise" (p) logically follow (*entailment*) or unfollow (*contradiction*) or are undetermined (*neutral*) to each other [34]. In other words:

Entailment: h is definitely true given p
Contradiction: h is definitely not true given p
Neutral: h might be true given p

To perform the prediction, an alignment was made between the *compatible*, *contradictory*, and *unrelated* classes of the ES_Headline_Contradiction dataset and the *entailment*, *contradiction*, and *neutral* classes annotated in the NLI datasets.

In this sense, the *compatible* and *contradictory* classes of our dataset correspond semantically to the *entailment* and *contradiction* classes of the NLI task dataset. The *entailment* between the headline and the body text indicates compatibility between them, whereas *contradiction* between headline and body text is exactly the same as our contradictory information definition.

However, the *unrelated* class of our dataset does not correspond exactly to the *neutral* class on the NLI task. In NLI, the *neutral* class indicates that the relation is undetermined (might be true or not), so in this case it is not known if there is a relation or not between both pieces of text. By contrast, in the domain of misleading headlines, we are interested in determining for the *unrelated* class that there is no relation at all between the headline and the body text, so it does not correspond exactly to the *neutral* class of the NLI. Therefore, these two classes cannot be directly aligned. In the case of this experiment, only the classes that have a direct correspondence with our classification are used, since we are obtaining the performance of the pre-trained models

TABLE 12. Results obtained in experiment 2: Detecting the types of contradiction.

Systems	F_1 Score (%) Neg	Ant	Num	Str	Fac	$F_1m(\%)$
BETO (*fine-tuned*)	94.21	93.49	**97.53**	94.85	69.04	89.82
RoBERTa-base-bne (*fine-tuned*)	**96.95**	**96.02**	97.52	96.23	74.86	92.32
Spanish_NLI_model (*fine-tuned*)	94.05	91.99	96.20	89.89	52.34	84.89
Multilingual_NLI_model (*fine-tuned*)	96.49	**96.00**	97.47	**96.60**	**76.19**	**92.55**

TABLE 13. Results obtained in experiment 3: Pre-trained NLI models vs fine-tuned models on our dataset to detect *compatible* and *contradictory* examples.

Systems	F_1 Score (%) Compatible	Contradictory	$F_1m(\%)$
Spanish_NLI_model (*pre-trained*)	20.43	**67.84**	44.14
Multilingual_NLI_model (*pre-trained*)	**68.18**	62.67	**65.42**
Spanish_NLI_model (*fine-tuned*)	86.64	86.04	86.34
Multilingual_NLI_model (*fine-tuned*)	**94.52**	**94.61**	**94.57**

in order to compare them with the improvement by using fine-tuning. Therefore, only the classes for which the system was pre-trained can be used and for this reason it was decided to ignore the *unrelated* class from ES_Headline_Contradiction and the *neutral* class from the predictions of the chosen models.

Two predictions were performed using each chosen model. First, a prediction was made with the pre-trained models, and then fine-tuning was performed on the training set of the ES_Headline_Contradiction dataset to compare the performance of each model before and after fine-tuning. Table 13 shows the results of these systems.

The pre-trained models on the NLI task obtained relatively discrete results to predict *compatible* and *contradictory* classes. In the case of the specific language model — Spanish_NLI_model (*pre-trained*)—, results below 50% of F_1 metric were obtained, worse than expected by a hypothetical system that predicts randomly this binary classification for the ES_Headline_Contradiction dataset. The poor performance of the models pre-trained for the NLI task to predict our dataset may be due to the word count difference between the texts annotated for our task and those annotated for the NLI task.

After performing the fine-tuning on the ES_Headline _Contradiction dataset, we obtained the Spanish_NLI_model (*fine-tuned*) and Multilingual_NLI_model (*fine-tuned*) systems. The results of the F_1m metric improved considerably, corresponding to results obtained in the experiments predicting all classes (section VI-A). These results demonstrate that for this task, fine-tuning on ES_Headline_Contradiction is the best decision compared to discrete results obtained by off-the-shelf NLI models. However, these models are a good alternative to create baseline systems after undergoing fine-tuning because of the similarity between tasks. Compared with the state-of-the-art system trained for the NLI task in English, the results are very similar, as in the case of [71] whose experiments delivered results of 93.1% accuracy in the SNLI benchmark, compared to 94.57% obtained by our best system to predict ES_Headline_Contradiction.

In this experiment, the Multilingual_NLI_model (*fine-tuned*) system also obtained the best results, outperforming the Spanish_NLI_model (*fine-tuned*) system by 8.23%. Both systems obtained similar results in both predicted classes, due to the quality of the training examples and the balanced number of examples of each class in this dataset. As indicated in the discussion of the first experiment, the results for class prediction could be improved by introducing external semantic information, similar to the introduction of Semantic Role Labelling [72] and the use of Wordnet relations [73], which can improve the results of deep learning models.

Finally, with the aim of determining which types of contradictions are the most complex to classify, a statistical analysis of the errors made by the model is carried out, classifying the *contradictory* examples. The Multilingual_NLI_model (*fine-tuned*) system was chosen, which obtained the best results classifying in *compatible* and *contradictory* examples. Figure 5 shows a bar chart by contradiction type. Each bar represents one of the five types of labels; in blue is the percentage of those classified correctly, and in red those classified incorrectly.

In line with the results obtained in section VI-B, the type of contradiction that obtains the worst results is the *factive* one. The figure shows that close to 50% of this type of contradiction is classified as *compatible*, which shows the need to increase the examples of this type of contradiction. In absolute terms, it is followed by the *structure* type, but as this type has the largest number of examples annotated, in percentage terms the result is similar to the rest of the types.

After carrying out a more in-depth qualitative analysis, we found that the model still makes mistakes in classifying contradictions of low complexity, such as the *numeric/date* or the *negation* ones. Regarding the *numeric/date* contradiction, there are two common mistakes. The first one is related to headlines containing numbers that do not appear in the body

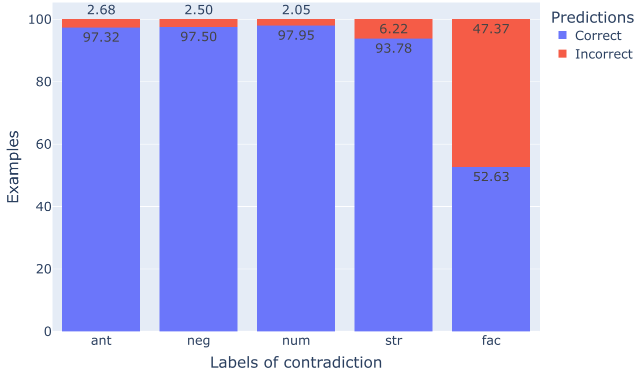

FIGURE 5. **Correct and incorrect predictions by contradiction type (in percentage terms).**

explicitly, as the text presents the elements as a list (i.e. *10 tips scammers don't want you to know this Black Friday*). The second one takes place when the number appearing in the headline which causes the contradiction also appears in the body text, even when in another context, leading to misclassification errors. Concerning the *negation* contradiction, the common misclassifications are: i) verbs that are negated both in the headline and in the body text, even if they are different verbs and ii) negative compound verbs that are considered semantically more complex.

Regarding the *antonyms* contradiction, the most common misclassification occurred in cases where there is a verb in the headline and that verb is nominalised in the body text. According to the UNE 153101 EX guideline[25], nominalisations in texts reduce understandability and in our case, this linguistic characteristic also confuses the model.

As can be seen in figure 5, the most complex contradiction types to predict are the *structure* and the *factive* ones. In the case of the *structure* contradiction, it has been noted that the presence of several named entities in headlines and body text negatively influence the correct prediction. Finally, concerning the *factive* contradiction, even if the main errors are observed in headlines or body text with compound verbs, we consider that the model requires a higher volume of training examples due to the high semantic comprehension needed to predict.

D. COMPARISON BETWEEN DATASETS VERSIONS

This experiment aims to evaluate the performance of the best language model used —Multilingual_NLI_model (*fine-tuned*)— based on the size of the different versions of the annotated datasets. The two preliminary versions of the dataset —*Dataset_v1* (obtained after the first annotation phase) and *Dataset_v2* (obtained after the second annotation phase)— and then the whole version —*ES_Headline_Contradiction* (merging Dataset_v1 and Dataset_v2)— were used to carry out this experiment.

When comparing the results between each version of the dataset (see table 14), the global F_1m of ES_Headline_Contradiction is slightly better than the second version of the dataset and also better than the first version of the dataset, where *factive* contradictions were not included. In terms of the *structure* contradiction, the F_1 metric indicates a significant increase in comparison with the first version of the dataset. This is likely to be due to the higher number of annotated examples made possible by the semi-automation in the dataset generation methodology. Indeed, the number of examples increased from 110 to 2,786, representing a 25 fold increase compared to the first version of the dataset.

In the case of *factive* contradictions, the lowest F_1 is obtained in comparison with the results obtained for the other types of contradictions, which was expected because only 6 percent of all the generated examples of contradictions correspond to this type. Thus, there is room for improvement in performance terms which requires an increase in the examples for this contradiction type.

E. LIMITATIONS OF THE PROPOSAL AND POTENTIAL SOLUTIONS

Although the results obtained in the application of the dataset are very satisfactory, the proposal has a number of limitations that should be addressed.

The first limitation is that the number of examples of *negation* and *factive* is very limited, especially the factives, causing a significant imbalance with respect to the other types of contradiction. This suggests that it would be necessary to expand this dataset so as to include a greater number of these types, aiming for their automatic generation, as in the case of other types of contradictions. As their automatic generation is more complicated, in the future we will consider the application of a model based on the human in the loop concept. In such a system, the modification for these two types of contradictions is self-generated, and through the intervention of the human, these examples are corrected or modified if necessary, and once revised, they are sent again to the training dataset. By this means the machine-human-machine interaction is generated.

Another limitation of the dataset in its current version is that the examples have been created and in the future we will consider introducing examples of real cases that are being used. This process is not simple because it is necessary to find headlines that are within the established types, and we will even have to consider the possibility of contemplating other types of contradictions present in a real scenario. In order to propose an efficient way to obtain the most relevant news for the training dataset, Active Learning techniques will be applied, so that with fewer examples the system will be able to learn without sacrificing the performance of the classifier.

Finally, at this point, the system is limited to make the contradiction classification decision based only on the semantics of the relationship between the headline and the news body text. In a real scenario, this solution should lead to a hybrid process, which does not only consider the content but also the

[25] https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060036

TABLE 14. Results obtained in experiment 4: Comparison between datasets versions.

	F_1 Score (%)					$F_1m(\%)$
Systems	**Neg**	**Ant**	**Num**	**Str**	**Fac**	
Dataset_v1	**97.91**	94.23	92.64	66.66	-	87.86
Dataset_v2	91.78	95.43	97.46	**97.46**	**78.96**	92.22
ES_Headline_Contradiction	96.49	**96.00**	**97.47**	96.60	76.19	**92.55**

context, where external knowledge sources are consulted in order to detect the contradictions known as World Knowledge (WK) contradictions [13], which is an area of research beyond the scope of this paper.

All these limitations will be addressed in future research.

VII. CONCLUSION

This research addresses misleading headlines detection through automatic contradiction identification. For this purpose, a novel methodology for building semi-automatic datasets is designed to enable us to create a new Spanish dataset (ES_Headline_Contradiction), in an efficient and effective way. This dataset annotates the semantic relationship between headlines and body text within three main categories: *compatible*, *contradictory* and *unrelated*. Furthermore, it also contains a fine-grained annotation that distinguishes the type of contradiction according to its characteristics, representing a novel contribution compared to the rest of the datasets relevant to this task. Five types of contradictions are covered, representing a broad spectrum of the contradictions defined by [13].

The semi-automatic process enabled the annotation of 25,945 news items, and approximately 18,000 of them were automatically annotated. The validation performed shows that both the construction process and the resulting dataset exhibit the quality required by the annotation guideline, suggesting that the potential exists for reproducing the process for other automatic annotation scenarios.

The results obtained by the experiments show that the created dataset is a good option to train models that accurately detect contradictions in Spanish (Best F1 of 95.63%) and therefore to support the identification of misleading headlines detection. Moreover, the dataset enables a more fine-grained detection of the type of contradiction with high accuracy (Best F1 92.55%). This can serve to enhance explainability in relation to information quality so as to assist journalists, fact-checkers and other users. Additionally, the experiments demonstrated that pre-trained language models constitute a viable option for the construction of baselines with a fine-tuning on the task dataset.

In future research, as stated in the Limitations subsection, we propose to extend the dataset following the automatic annotation approach by downloading news from other reliable and unreliable data sources so that the diversity of news sources is addressed. Headlines extracted from real-world media with contradictory news items would open the spectrum for better coverage of misleading headlines. Furthermore, the methodology could be applied to other domains, apart from the political and economic ones. More examples of the contradiction type *factive* will be included, as well as the rest of the contradictions defined by [13]. Experiments will be performed to determine how the overlapping between contradictory categories impacts the contradiction task detection and how the degree of headline modifications affects the result of classification.

In addition, it would be of interest to enhance the contradiction detection by enriching the process with external information resources, which would enable training to be done efficiently on fewer examples. This would lead to a more robust approach to detecting misleading headlines, and it could be very effective for headlines produced in a real world scenario. Moreover, annotating if a headline is intentionally deceptive would enable an evaluation of the impact of contradiction recognition in the task of deception detection. Finally, future applications of this proposal could involve integrating the contradiction detection technology into user content applications as a plugin or chat bot, warning users whenever a headline and news body text present contradictory information. Although content providers may not initially be interested in incorporating this tool, once they acknowledge the benefits, we expect the uptake to be strong. This is because integrating the contradiction detection technology proposed into the content providers platforms will serve to indicate a quality standard attained for information reliability. Moreover, as the cost and effort of producing disinformation declines, with the spread of publicly available language generation models, many news items will be created by machines. This potentially increases the risk of an even greater dissemination of poor quality and unreliable information. Incorporating these types of tools that guarantee the quality and accuracy of information will become a necessity for news providers. In essence, the market will eventually judge content providers that do not have some type of disinformation detection accordingly, i.e perhaps with some suspicion and doubt.

REFERENCES

[1] A. T. van Dijk, *News as Discourse* (Communication Series). Lawrence Erlbaum Associates, 1988.

[2] J. Reis, F. Benevenuto, O. S. P. V. de Melo, R. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," in *Proc. 9th Int. Conf. Web Social Media*, 2015, pp. 357–366.

[3] J. Kuiken, A. Schuth, M. Spitters, and M. Marx, "Effective headlines of newspaper articles in a digital environment," *Digit. Journalism*, vol. 5, no. 10, pp. 1300–1314, Nov. 2017.

[4] D. Dor, "On newspaper headlines as relevance optimizers," *J. Pragmatics*, vol. 35, no. 5, pp. 695–721, May 2003.

[5] C. E. J. Normala, I. Iskandar, F. Sidi, and A. L. Suriani, "Fakeheader: A tool to detect deceptive online news based on misleading news headlines and contents," *Turkish J. Comput. Math. Educ. (TURCOMAT)*, vol. 12, no. 3, pp. 2217–2223, Apr. 2021.

[6] U. K. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai, "The effects of subtle misinformation in news headlines," *J. Exp. Psychol., Appl.*, vol. 20, no. 4, p. 323, 2014.

[7] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on Twitter?" *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 44, no. 1, pp. 179–192, Jun. 2016.

[8] W. Wei and X. Wan, "Learning to identify ambiguous and misleading news headlines," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4172–4178.

[9] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as 'false news,'" in *Proc. ACM Workshop Multimodal Deception Detection*, Nov. 2015, pp. 15–19.

[10] Y. Chen, J. N. Conroy, and L. V. Rubin, "News in an online world: The need for an 'automatic crap detector,'" in *Proc. 78th ASIS Annu. Meeting, Inf. Sci. Impact, Res. Community*, 2015, pp. 1–4.

[11] S. Chesney, M. Liakata, M. Poesio, and M. Purver, "Incongruent headlines: Yet another way to mislead your readers," in *Proc. EMNLP Workshop, Natural Lang. Process. Meets Journalism*, 2017, pp. 56–61.

[12] K. Shu, S. Wang, D. Lee, and H. Liu, *Disinformation, Misinformation, and Fake News in Social Media*. Berlin, Germany: Springer, 2020.

[13] M.-C. D. Marneffe, A. N. Rafferty, and C. D. Manning, "Finding contradictions in text," in *Proc. Assoc. Comput. Linguistics*, 2008, pp. 1039–1047.

[14] R. M. Monarch, *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. New York, NY, USA: Simon and Schuster, 2021.

[15] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: A web-based tool for NLP-assisted text annotation," in *Proc. Demonstrations Session (EACL)*, ACL, Apr. 2012, pp. 102–107.

[16] R. Sepulveda-Torres, A. Bonet-Jover, and E. Saquete, "'Here are the rules: Ignore all rules': Automatic contradiction detection in Spanish," *Appl. Sci.*, vol. 11, no. 7, p. 3060, Mar. 2021.

[17] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats and potential mitigations," 2023, *arXiv:2301.04246*.

[18] B. Pouliquen, R. Steinberger, and C. Best, "Automatic detection of quotations in multilingual news," in *Proc. Recent Adv. Natural Lang. Process.*, Borovets, Bulgaria, 2007, pp. 487–492.

[19] A. Vlachos and S. Riedel, "Identification and verification of simple claims about statistical properties," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 2596–2601.

[20] J. Thorne and A. Vlachos, "An extensible framework for verification of numerical claims," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 37–40.

[21] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer: A parse-and-trim approach to headline generation," in *Proc. HLT-NAACL Text Summarization Workshop*, 2003, pp. 1–8.

[22] C. Stab and I. Gurevych, "Recognizing insufficiently supported arguments in argumentative essays," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, Volume 1, Long Papers*, Valencia, Spain, 2017, pp. 980–990.

[23] G. Zarrella and A. Marsh, "MITRE at SemEval-2016 task 6: Transfer learning for stance detection," in *Proc. 10th Int. Workshop Semantic Eval.*, 2016, pp. 458–463.

[24] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance detection in web and social media: A comparative study," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Cham, Switzerland: Springer, 2019, pp. 75–87.

[25] W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, San Diego, CA, USA, 2016, pp. 1163–1168.

[26] M. Babakar, N. Bakos, H. Daumé, A. Mantzarlis, D. Seddah, A. Vlachos, and C. Wardle. (2016). *Fake News Challenge-I*. accessed: Jan. 21, 2021. [Online]. Available: http://www.fakenewschallenge.org/

[27] Q. Zhang, S. Liang, A. Lipani, Z. Ren, and E. Yilmaz, "From stances' imbalance to their hierarchical representation and detection," in *Proc. World Wide Web Conf.*, May 2019, pp. 2323–2332.

[28] C. Dulhanty, L. Jason Deglint, I. B. Daya, and A. Wong, "Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection," 2019, *arXiv:1911.11951*.

[29] R. Sepulveda-Torres, M. Vicente, E. Saquete, E. Lloret, and M. Palomar, "HeadlineStanceChecker: Exploiting summarization to detect headline disinformation," *J. Web Semantics*, vol. 71, Nov. 2021, Art. no. 100660.

[30] S. Yoon, K. Park, J. Shin, H. Lim, S. Won, M. Cha, and K. Jung, "Detecting incongruity between news headline and body text via a deep hierarchical encoder," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell.*, 2019, pp. 791–800.

[31] O. Levy, T. Zesch, I. Dagan, and I. Gurevych, "Recognizing partial textual entailment," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 451–455.

[32] S. Harabagiu, A. Hickl, and F. Lacatusu, "Negation, contrast and contradiction in text processing," in *Proc. 21st Nat. Conf. Artif. Intell.*, 2006, pp. 755–762.

[33] R. Sifa, M. Pielka, R. Ramamurthy, A. Ladi, L. Hillebrand, and C. Bauckhage, "Towards contradiction detection in German: A translation-driven approach," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 2497–2505.

[34] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 632–642.

[35] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 1112–1122.

[36] A. Conneau, G. Lample, R. Rinott, A. Williams, R. S. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating cross-lingual sentence representations," 2018, *arXiv:1809.05053*.

[37] Y. Takabatake, H. Morita, D. Kawahara, S. Kurohashi, R. Higashinaka, and Y. Matsuo, "Classification and acquisition of contradictory event pairs using crowdsourcing," in *Proc. 3rd Workshop EVENTS, Definition, Detection, Coreference, Represent.*, Denver, Colorado, 2015, pp. 99–107.

[38] Z. Rahimi and M. Shamsfard, "Contradiction detection in Persian text," 2021, *arXiv:2107.01987*.

[39] E. Saquete, D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar, "Fighting post-truth using natural language processing: A review and open challenges," *Exp. Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112943.

[40] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, and J. J. M. Escobar, "Detection of fake news in a new corpus for the Spanish language," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4869–4876, May 2019.

[41] M. Evrard, R. Uro, N. Herve, and B. Mazoyer, "French tweet corpus for automatic stance detection," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 6317–6322.

[42] A. B. Abacha, D. Dinh, and Y. Mrabet, "Semantic analysis and automatic corpus construction for entailment recognition in medical texts," in *Proc. Conf. Artif. Intell. Med. Eur.* Cham, Switzerland: Springer, 2015, pp. 238–242.

[43] G. K. Shahi and D. Nandini, "FakeCOVID—A multilingual cross-domain fact check news dataset for COVID-19," in *Proc. 14th Int. AAAI Conf. Web Social Media*, S. Chancellor, K. Garimella, and K. Weller, Eds. Atlanta, GA, USA, 2020, pp. 1–9.

[44] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annu. Meeting Assoc. for Comput. Linguistics*, R. Barzilay, and M. Kan, Eds. Vancouver, BC, Canada, 2017, pp. 422–426.

[45] T. Mitra and E. Gilbert, "CREDBANK: A large-scale social media corpus with associated credibility annotations," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 9, 2015, pp. 258–267.

[46] M. Färber, V. Burkard, A. Jatowt, and S. Lim, "A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2020, pp. 3007–3014.

[47] V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1120–1125.

[48] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: A study of annotation selection criteria," in *Proc. NAACL HLT Workshop Act. Learn. Natural Lang. Process.*, 2009, pp. 27–35.

[49] P. Liu, Y. Deng, C. Zhu, and H. Hu, "XCMRC: Evaluating cross-lingual machine reading comprehension," in *Natural Language Processing and Chinese Computing*. Dunhuang, China: Springer, 2019, pp. 552–564.

[50] C. Bucaria, "Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines," *Humor-Int. J. Humor Res.*, vol. 17, no. 3, pp. 279–309, Jan. 2004.

[51] Y. M. Salih and Q. Abdulla, "Linguistic features of newspaper headlines," *J. Al-Anbar Univ. Lang. Literature*, vol. 7, pp. 192–213, Jan. 2012.

[52] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7315–7330.

[53] V. Lingam, S. Bhuria, M. Nair, D. Gurpreetsingh, A. Goyal, and A. Sureka, "Deep learning for conflicting statements detection in text," *PeerJ Prepr.*, vol. 6, p. e26589, Mar. 2018.

[54] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo, "The timebank corpus," in *Corpus Linguistics*. Lancaster, U.K.: University Centre for Computer Corpus Research on Language, University of Lancaster, 2003, p. 40.

[55] J. Surm, "AFP, EFE and DPA as international news agencies," *Journalism*, vol. 21, no. 12, pp. 1859–1876, Dec. 2020.

[56] M. V. G. Clavero, "Agencias de noticias, su constante reinvención como estrategia para enfrentar la competencia," *Estudios Sobre el Mensaje Periodístico*, vol. 22, no. 1, pp. 329–341, May 2016.

[57] A. Hanselowski, A. Pvs, A. B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A retrospective analysis of the fake news challenge stance-detection task," in *Proc. 27th Int. Conf. Comput. Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, NM, USA, Aug. 2018, pp. 1859–1874.

[58] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[61] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in *Proc. 20th Chin. Nat. Conf. Comput. Linguistics*, Huhhot, China, Aug. 2021, pp. 1218–1227.

[62] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, Istanbul, Turkey, May 2012, pp. 2214–2218.

[63] J. Canete, G. Chaperon, R. Fuentes, and J. Perez, "Spanish pre-trained BERT model and evaluation data," in *Proc. ICLR*, 2020, pp. 1–10.

[64] A. G. Fandiño, J. A. Estapé, M. Pamies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas, "MarIA: Spanish language models," *Procesamiento del Lenguaje Natural*, vol. 68, pp. 39–60, Mar. 2022.

[65] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–23.

[66] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A new benchmark for natural language understanding," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4885–4901.

[67] A. Liu, S. Swayamdipta, A. N. Smith, and Y. Choi, "WANLI: Worker and AI collaboration for natural language inference dataset creation," in *Findings of the Association for Computational Linguistics: EMNLP*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6826–6847.

[68] A. Parrish, W. Huang, O. Agha, S. Lee, N. Nangia, A. Warstadt, K. Aggarwal, E. Allaway, T. Linzen, and R. S. Bowman, "Does putting a linguist in the loop improve NLU data collection?" in *Findings of the Association for Computational Linguistics: EMNLP*, M. Moens, X. Huang, L. Specia, and S. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4886–4901.

[69] M. Laurer, W. V. Atteveldt, A. Casas, and K. Welbers, "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI," *Political Anal.*, pp. 1–33, Jun. 2022.

[70] X. Kang, B. Li, H. Yao, Q. Liang, S. Li, J. Gong, and X. Li, "Incorporating synonym for lexical sememe prediction: An attention-based model," *Appl. Sci.*, vol. 10, no. 17, p. 5996, Aug. 2020.

[71] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, "Entailment as few-shot learner," 2021, *arXiv:2104.14690*.

[72] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," in *Proc. 34th AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, pp. 9628–9635.

[73] Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, and S. Wei, "Neural natural language inference models enhanced with external knowledge," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2406–2417.

ROBIERT SEPÚLVEDA-TORRES received the M.Sc. degree in applied informatics from the Technological University of Havana, Cuba, in 2018, and the Ph.D. degree in informatics from the University of Alicante, Spain, in 2022. He is currently a Postdoctoral Researcher with the Department of Software and Computing Systems, University of Alicante, applying natural language processing and machine learning techniques to solve problems related to disinformation and credibility. He has participated as a researcher in four research projects funded by public calls.

ALBA BONET-JOVER received the degree in translation and interpretation from the University of Alicante, Spain, in 2018, the double master's degree in institutional, commercial, and legal translation from the University of Lyon, France, in 2019, and the Ph.D. degree in human language technologies from the University of Alicante, in 2023, with a focus on the reliability modeling in news through language and NLP. She works as a Postdoctoral Researcher with the Department of Software and Computing Systems, University of Alicante. She has published six papers and a book chapter.

ESTELA SAQUETE has been a Professor with the University of Alicante, since 2002 and belongs to the Research Group "Processing of Language and Information Systems." As a result of all her research activity, a total of 92 publications are highlighted. Of these, 59 publications are contributions in high-impact journals, indexed in JCR or with equivalent quality criteria and 33 papers are contributions to prestigious international research conferences in the field of natural language processing. She has an H-index of 8 (according to WOS), 11 (according to Scopus), and 17 (according to Google Scholar). She has participated as a researcher in 49 R&D&I projects financed with European, national, and regional public funds (four of them as the project manager) and ten R&D&I contracts (not competitive), agreements, or projects with companies and/or governments (two of them as the project manager). She is currently the principal investigator (PI1) of two national R&D&I plan projects, which have the main objective of studying and developing different human language technology (HTL) techniques, resources, and tools directed toward modeling digital entities, their relationships in social media, and their evolution over time.